Sparse-H3D—Transformer-based Human3D Segmentation Using Sparse Depth

Swastik Mahapatra [1], Anirudh Srihari [2], Parth Gupta [3], Shreya Shri Ragi [4]

***Abstract*— Robust 3D human segmentation is crucial for the safe operation of mobile robots, such as autonomous vehicles and drones, in real-world environments. However, state-of-the-art segmentation models like Human3D are predominantly trained on dense, synthetic indoor datasets and struggle to generalize to outdoor settings, where robots rely on sparse depth data from sensors like LiDAR. In this work, we present Sparse-H3D, a framework that adapts transformer-based 3D human segmentation to operate effectively on sparse outdoor point clouds. Our approach combines two strategies: (1) fine-tuning a Human3D transformer model on LiDAR-style downsampled data to enhance its performance on sparse inputs, and (2) developing preprocessing pipelines to upsample and align sparse outdoor LiDAR data, making it more compatible with existing dense-data-trained models. Experiments across both synthetic (Egobody) and real-world (SynLiDAR) datasets demonstrate that these methods significantly improve segmentation accuracy in sparse, outdoor scenarios. Ablation studies further highlight the importance of input density and orientation alignment, and our results show that domain-specific fine-tuning, together with intelligent preprocessing, can bridge the gap between indoor-trained models and real-world outdoor applications.**

I. INTRODUCTION

Robust 3D human segmentation is essential for enabling mobile robot such as autonomous vehicles and drones to operate safely in real-world environments. Existing state-of-the-art models, like Human3D [**human3d**], are primarily trained on dense, synthetic indoor data and struggle to generalize to outdoor settings, where robots must rely on sparse depth information from sensors such as LiDAR. This project, Sparse-H3D, aims to address these limitations by adopting two approaches. First, fine-tuning a transformer-based segmentation network for human detection using sparse depth data, and second, designing pre-processing methods to reduce sparsity in outdoor Lidar data and allow existing Human3D models to segment humans.

[1] Swastik Mahapatra is a Master's Student in Robotic Systems Development at Robotics Institute, SCS, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2] Anirudh Srihari is a Master's Student in Robotic Systems Development at Robotics Institute, SCS, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[3] Parth Gupta is a Master's Student in Robotic Systems Development at Robotics Institute, SCS, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[4] Shreya Shri Ragi is a Master's Student in Robotic Systems Development at Robotics Institute, SCS, Carnegie Mellon University, Pittsburgh, PA 15213, USA

II. MOTIVATION

First responders face challenges with visibility when navigating smoke-filled environments during emergencies like fires or industrial accidents. But, traditional RGB-based perception systems fail in these conditions, creating an urgent need for thermal imaging and RaDAR-based solutions that can operate through perceptual degradation. As part of our CMU AirLab capstone project, we developed an autonomous drone system to address this gap, using stereo thermal cameras to generate 3D maps in perceptually degraded scenarios.

The key technical hurdle lies in adapting human segmentation models to sparse point clouds from thermal disparity calculations – data fundamentally different from the dense Depth inputs used by state-of-the-art models like Human3D. While Human3D's synthetic training pipeline (which simulates Kinect sensor noise via SimKinect [**SimKinect**]) shows the viability of sensor-aware domain adaptation, no equivalent work exists for the sparse 3D data produced by our first-responder drone sensor.

Our work bridges this gap through two strategies:

1) **Sensor simulation for training:** Inspired by Human3D's SimKinect, we developed preprocessing pipelines to convert dense synthetic data into sparse thermal/LiDAR-like point clouds, enabling targeted model fine-tuning.

2) **Real-data enhancement:** For field deployment, we designed upsampling and alignment techniques to improve segmentation reliability on raw sensor outputs, as demonstrated by mmParse's success [**wang2023human**] in parsing sparse mmWave radar data through smoke.

This dual approach ensures our system can localize humans while leveraging the learnings from existing HUman3D model, while providing a blueprint for adapting indoor-trained models to outdoor robotic platforms. By closing the sparsity structure gap for sparse thermal/LiDAR data, we aim to enhance situational awareness in disaster scenarios where conventional vision systems fail.

Fig. 1. First responder drone

III. RELATED WORK

Human Segmentation in Point Clouds Human3D is a pioneering work in the field of 3D human segmentation, which introduced the first multi-human body segmentation model that operates directly on 3D scenes. This model addresses the challenge of segmenting humans in cluttered 3D environments by leveraging synthetic data for pre-training, a task that has been limited by the scarcity of annotated training data for humans interacting with 3D scenes. But until now, the efficacy of this approach has only been proven for indoor environments using the Egobody [**zhang2022egobody**] dataset.

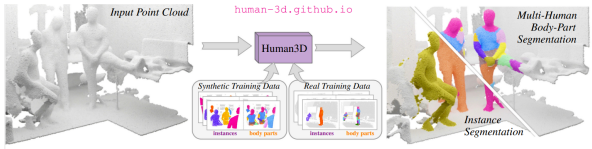

Fig. 2. Source: https://human-3d.github.io/assets/Human3D_paper.pdf

Outdoor 3D Human Segmentation In the context of outdoor data, work on 3D human body segmentation has been very limited. For instance, papers like LidPose [**zhang2022egobody**], present a vision-transformer based method for real-time human skeleton estimation in sparse LiDAR point clouds. This approach transforms NRCS LiDAR point cloud data into a 2D representation, enabling skeletal estimation for pose detection. While this is good for real-time pose detection applications, it does not help with accurate point cloud segmentation.

Datasets of LiDAR Point Clouds with Humans in Outdoor Setting Recent developments with datasets like SLOPER4D [**Dai˙2023˙CVPR**], equips us with relevant outdoor human point cloud data with annotations to train good segmentation models. Older datasets like nuScenes [**nuscenes**] may also help with additional data if necessary.

Fig. 3. Source: http://www.lidarhumanmotion.net/sloper4d/

Fig. 4. Source: https://www.nuscenes.org/nuscenes

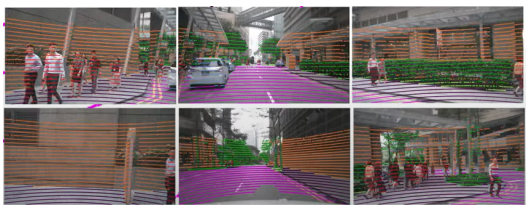

Fig. 5. Source: https://www.nuscenes.org/nuscenes

SynLiDAR [**xiao2022transfer**] is a large-scale, high-fidelity synthetic dataset specifically designed for autonomous driving scenarios and includes dense point clouds with rich semantic annotations. While SynLiDAR itself is not human-centric, its high density and diversity in scenes make it a useful candidate for pretraining or augmenting human segmentation models in outdoor contexts. In our approach, we propose to upsample point clouds from SynLiDAR using point cloud processing techniques (e.g., interpolation or learning-based super-resolution) to increase density, and then run inference using the original Human3D model, which was trained on dense indoor depth maps. This allows us to evaluate the generalizability of Human3D in outdoor scenarios with synthetic yet dense LiDAR data.

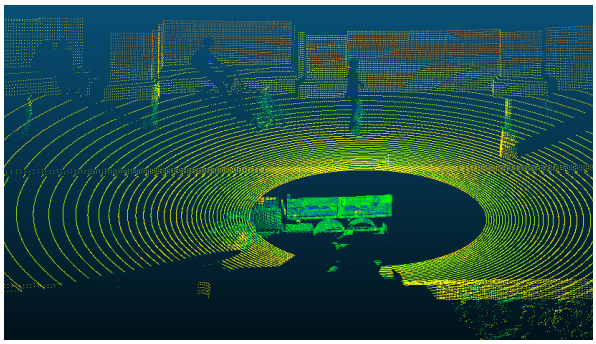

Fig. 6. Sample Point Cloud from SynLidar

Our work builds upon these advancements, particularly leveraging the model performance gained with the synthetic data generation approach of Human3D, while focusing specifically on sparse outdoor data. We aim to adapt and extend existing models to better handle the unique challenges presented by large-scale outdoor point clouds, such as varying densities, occlusions, and diverse human poses in complex environments.

IV. METHODOLOGY

A. Preprocessing

To adapt existing 3D segmentation models trained on dense indoor point clouds to the sparse and noisy nature of outdoor LiDAR data, we developed several pre-processing techniques. These serve two main purposes: (1) simulate LiDAR sparsity on dense indoor datasets to fine-tune models appropriately, and (2) enhance sparse outdoor LiDAR data to improve segmentation accuracy.

Fig. 7. Original Ego body dataset

1) For Training: Pre-processing Egobody dataset to simulate LiDAR-like behavior:

a) Voxel-Based Downsampling: For simulating sparse input from high-density indoor data, we used voxel grid filtering. The input point cloud is divided into a regular 3D voxel grid, and one representative point is selected per occupied voxel. This method reduces point density uniformly and emulates reduced resolution while maintaining overall scene structure.

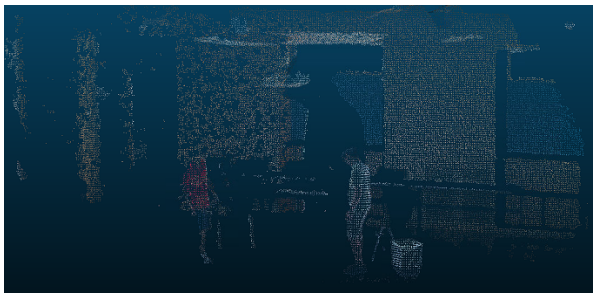

Fig. 8. Voxel-based Downsampling on Egobody

b) LiDAR-like Downsampling: To mimic the sampling pattern of a spinning LiDAR, we implemented a custom channel-based radial slicing algorithm. The method bins points based on their vertical height (y-axis) into a fixed number of channels, and further filters them by angular slices in the x-z plane (simulating azimuth bins). This approach models real LiDAR behavior more accurately than voxel grids and retains the sparsity and anisotropy characteristics of outdoor point clouds.

Fig. 9. Lidar-like Downsampling on Egobody

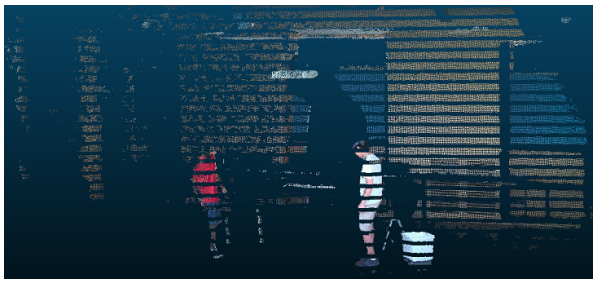

Fig. 10. Lidar-like Downsampling on Egobody

2) Upsampling Sparse LiDAR Data: To improve segmentation on inherently sparse outdoor LiDAR point clouds, we designed a lightweight upsampling technique. Starting from the original point cloud, we applied random per-point translations to generate two shifted copies of the data. These augmentations were stacked together with the original to increase point density. A statistical outlier removal filter was then applied to remove noise introduced during upsampling. Importantly, instead of applying a single global transformation, each point was randomly shifted independently, better simulating realistic perturbations while avoiding mode collapse in spatial features.

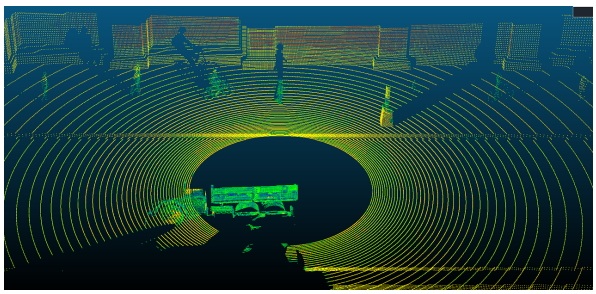

Fig. 11. Original SynLidar point cloud

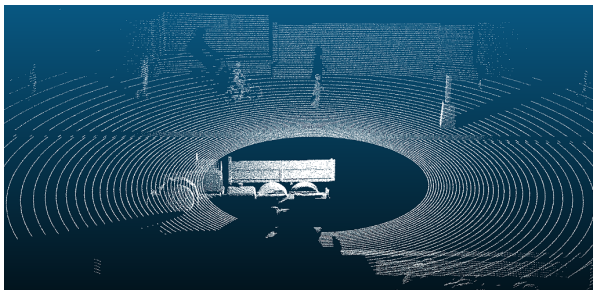

Fig. 12. Upsampled SynLidar point cloud

B. Fine-Tuning

To adapt the Human3D segmentation model to sparse, LiDAR-like outdoor point clouds, we fine-tuned the pre-trained transformer model provided by the original authors. The model was initialized using their publicly available checkpoints, which had been trained on dense synthetic indoor data from the Egobody dataset.

We then fine-tuned this model on the original Egobody dataset, after applying our custom LiDAR-like downsampling strategy to all training and validation samples. This downsampling simulated a 32-channel rotating LiDAR to better match the sparsity characteristics of outdoor point clouds.

Key aspects of the fine-tuning process included:

- **Checkpoint Initialization**: We used the pretrained Human3D weights as a starting point to leverage prior knowledge of human geometry and structure.
- **LiDAR-style Input**: All inputs during training were preprocessed using our LiDAR-like downsampling technique, limiting point cloud density while preserving vertical and azimuthal structure.
- **Positional Encoding Adaptation**: We retained the original transformer architecture and positional encodings, relying on the fine-tuning phase to implicitly adapt to the new spatial distribution.
- **Loss Function and Training**: The model was optimized using cross-entropy loss. We train the model for 23 epochs on randomly extracted 2000 point clouds while lowering the learning rate and freezing the first 20 layers.
- **Validation**: Model performance was monitored using a held-out subset of the down-sampled Egobody data to ensure generalization.

This process allowed the model to retain high-level geometric priors from dense data while adapting to the structured sparsity of LiDAR-like observations.

V. EXPERIMENTS

To evaluate the generalization capability of our model and understand the impact of different data preprocessing steps, we conducted a series of experiments using both the original and fine-tuned checkpoints of the pre-trained model from *Human3D*. Our experimental design focused on two datasets: the LiDAR outdoor *SynLidar* dataset and the indoor *EgoBody* dataset, with various preprocessing pipelines applied to both.

We began by performing inference using the original, unmodified checkpoint and different confidence thresholds, on the following data variants:

- **SynLidar (raw)**: The model was tested directly on the LiDAR data without any preprocessing.
- **EgoBody (voxel-downsampled)**: The input was voxel downsampled to mimic sparse LiDAR-like observations.
- **EgoBody (LiDAR-downsampled)**: We applied LiDAR-style depth-based downsampling to bring the dense EgoBody data closer to real-world LiDAR characteristics.

Next, to better match the EgoBody domain, we created a processed version of SynLidar by

- Upsampling the point cloud to decrease sparsity
- Cropping the original scene to extract a smaller room-sized scene with humans
- Rotating the point clouds about the global X-axis to match the orientation of the egobody dataset.
- Translating the point cloud such that the origin of the world coordinate system falls in the center of the point cloud.

These transformations helped us modify SynLidar point clouds to align with the distribution of point clouds in the Egobody dataset. We then ran inference and visually evaluated the performance of the original checkpoint on this processed SynLidar data.

To isolate the impact of each transformation step, we performed ablation experiments by removing one preprocessing operation (upsampling, cropping, rotation, translation) at a time. These experiments revealed that rotation alignment and upsampling are essential for model performance, while cropping and translation had negligible effects. This suggests that the original Human-3D model is not fully rotation invariant and requires dense data.

Finally, we fine-tuned the model on LiDAR-downsampled EgoBody data, initializing from the original checkpoint. We then inferred the fine-tuned model on:

- SynLidar (raw)
- LiDAR-downsampled EgoBody

These experiments were designed to test the model's adaptability to synthetic data, the influence of domain shift, and the effectiveness of downsampling and fine-tuning in closing the domain gap between synthetic and real-world data.

A. Original Model Inference on SynLidar data

We run the original model on SynLidar data to check the base model performance. It can be observed from the inference result image that the model was not able to segment the humans from the point cloud, indicating that the Lidar data is too sparse for the model.

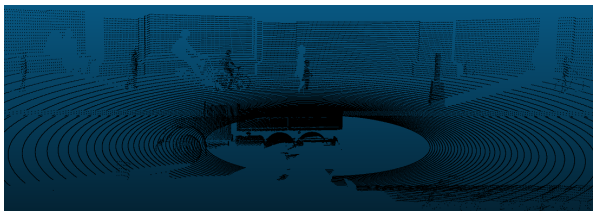

Fig. 13. Original Model Inference on SynLidar Data

B. Original Model Inference on Voxel Downsampled EgoBody data

We downsample the EgoBody dataset based on voxel methods. The model did not segment the humans since the point cloud is too sparse.

Fig. 14. Original Model Inference on Voxel Downsampled EgoBody data

C. Original Model Inference on Lidar downsampled Egobody data

We downsample the EgoBody dataset by retaining point clouds in channels. The model cannot segment the humans as the data is too sparse.

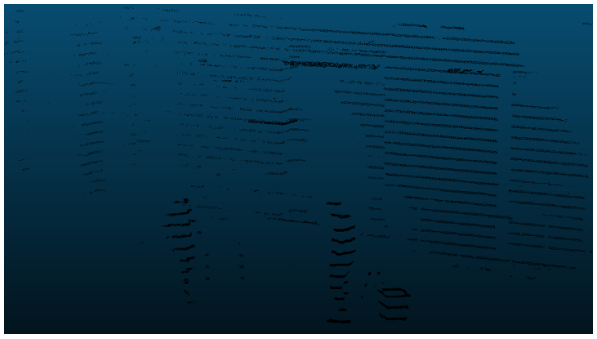

Fig. 15. Enter Caption

D. Original Model Inference on Rotation, Translation, Density and Scale Processed SynLidar Data

We processed the SynLidar Dataset to be closer to the EgoBody dataset by rotating, cropping, and increasing the density of the point cloud. The resulting point clouds is visualized in the following image.

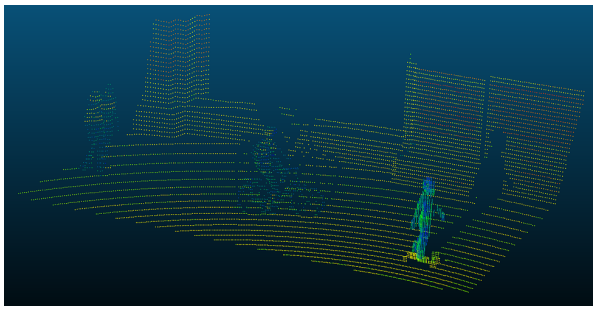

Fig. 16. Rotation, Translation, Density, and Scale processed SynLidar Data

We run the original Human-3D model on the new processed data, and we can see that the model is able to segment the human closest to the Lidar sensor.

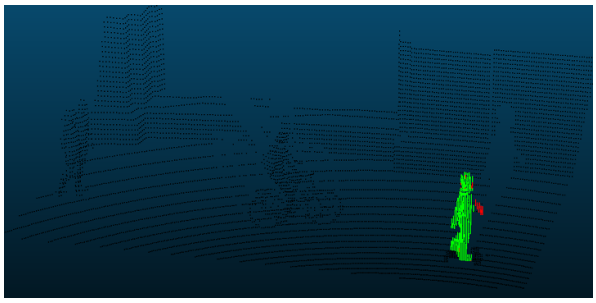

Fig. 17. Original Model Inference on Processed SynLidar Data

E. Finetuned Model Inference on SynLidar data

The Human3D model was unable to segment humans from the original point clouds of the SynLidar dataset. However, after fine-tuning, the new model is able to segment one humans with relatively higher density from the lidar point cloud, as seen in the following figure.

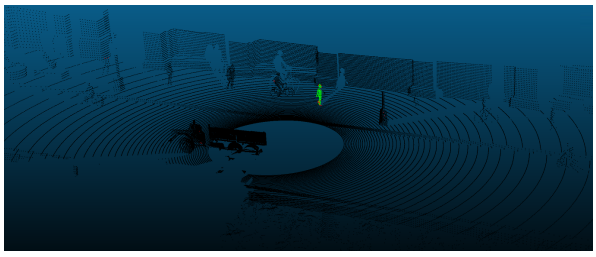

Fig. 18. Finetuned Model Inference on Original SynLidar Data

F. Finetuned Model Inference on Lidar downsampled Egobody data

When we previously ran the inference on a point cloud from egobody that has been downsampled using a lidar-like downsampling, we observed that the original human3d model was not able to segment any humans. However, after fine-tuning the model using downsampled point clouds, the model can segment humans successfully, even from sparse point clouds, as the one shown in the following figure.

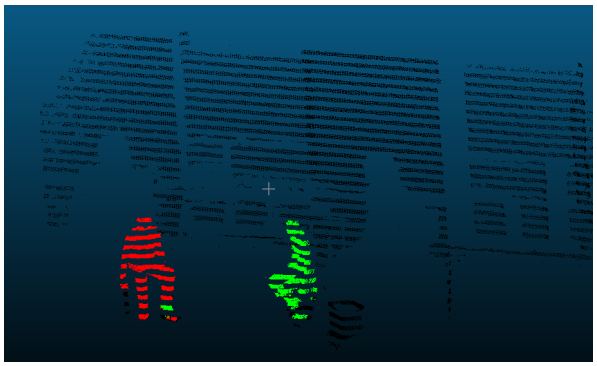

Fig. 19. Inference of fine-tuned model on lidar-like downsampling

G. Finetuned Model Inference on Upsampled SynLidar Point Cloud

We combine the two approaches - upsampling original LiDAR data to reduce sparsity and fine-tuning human3d on sparse data. The inference results from this pipeline are shown in the following figure. One human in the point cloud has been successfully segmented by the model.

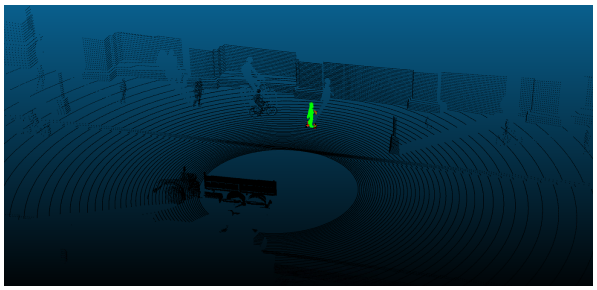

Fig. 20. Inference of fine-tuned model on up-sampled LiDAR data

VI. EVALUATION

We evaluated the Mask3D model using the author's official pretrained checkpoint on two datasets: EgoBody and a LiDAR downsampled version of EgoBody. Additionally, we also included performance results of our own fine-tuned checkpoints on the downsampled dataset to assess improvements.

A. Baseline Evaluation on EgoBody Dataset

On the high-resolution original dataset, the Mask3D checkpoint demonstrates excellent performance in semantic instance segmentation. Table I presents a summary of key evaluation metrics. The model achieves a high average precision (AP) across thresholds and strong intersection-over-union (IoU) values, indicating reliable predictions and accurate spatial localization. All loss terms remain low, confirming stable optimization.

B. Evaluation on LiDAR-like Dataset (32-Channels)

In contrast, when evaluated on the downsampled dataset, the model exhibits a dramatic performance degradation. The human segmentation AP drops by over 90%, and IoU for the human class falls below 10%. These results suggest that the pretrained Mask3D model relies heavily on the dense geometric structure of point clouds and generalizes poorly to sparse input data without retraining or domain adaptation.

C. Evaluation of Finetuned Model on Original Egobody Dataset

After fine-tuning the Human3d model on sparse point clouds, we evaluated the model on the original dense point clouds from the Egobody dataset. We observed that the Average Precision for human segmentation on the validation set drops from 92.14% to 60.0%.

D. Evaluation of Fine-Tuned Checkpoints on Downsampled Data (Our Method)

To address the limitations observed with the pretrained checkpoint, we fine-tuned the Mask3D model on the 32-channel downsampled dataset. The finetuning aimed to improve segmentation performance under sparse sensing conditions. The results of this retraining are included alongside the baseline metrics in Table I. We observed that the Average Precision for human segmentation on the validation set goes up from 10% to 90.0%.

TABLE I
PERFORMANCE OF MASK3D ON EGOBODY VS. DOWNSAMPLED EGOBODY (32-CHANNEL LIDAR-LIKE) VS. OUR METHOD

Metric	EgoBody	LiDAR EgoBody	Our Method
val_AP_human	0.9214	0.0077	0.9085
val_AP_25_human	0.9951	0.0547	0.9998
val_AP_50_human	0.9847	0.0217	0.9946
val_iou_human	0.9275	0.0879	0.9048
val_iou_background	0.9933	0.9188	0.9896
val_mean_iou	0.9604	0.5034	0.9472
val_loss_ce	0.0245	3.3387	0.0137
val_loss_dice	0.1030	1.7446	0.1246
val_loss_mask	0.0504	1.6828	0.0630

VII. CONCLUSION

A. Processing LiDAR Data to Leverage Human Segmentation Models Trained on Dense Point Clouds

In this work, we explored the feasibility of applying the Mask3D [**Schult23ICRA**] segmentation model—pre-trained on dense indoor point clouds—to sparse outdoor LiDAR data. Our experiments across the EgoBody and SynLiDAR datasets showed that the original model fails to generalize to sparse point clouds, producing poor or no segmentation results. To address this, we applied a series of preprocessing steps to the SynLiDAR data, including upsampling, rotation alignment, cropping, and translation.

Through ablation studies, we determined that upsampling the point cloud to increase density and rotating it to match the Egobody dataset's orientation significantly improved segmentation results, even without re-training the model. These findings highlight the model's sensitivity to input density and orientation, and emphasize that preprocessing plays a crucial role in adapting dense-data-trained models to sparse LiDAR settings.

B. Fine-tuning Human Segmentation Models on Sparse Point Clouds of Outdoor Settings

Beyond preprocessing, we investigated the effectiveness of fine-tuning the Human3D model on LiDAR-style sparse point clouds. Specifically, we fine-tuned the original checkpoint using a LiDAR-downsampled version of the EgoBody dataset and then evaluated performance on both raw SynLiDAR and downsampled EgoBody data.

Results show that fine-tuning improves the model's adaptability to sparse inputs, allowing it to partially recover segmentation capability on previously unsegmentable data. While performance still lags behind results on dense inputs, this approach demonstrates a promising path forward: synthetic sparse datasets like SynLiDAR, when combined with sparse-aware fine-tuning, can help bridge the domain gap and extend indoor-trained models to realistic outdoor environments.

Together, these two strategies—intelligent preprocessing and domain-specific fine-tuning—pave the way toward robust, real-world 3D human segmentation in mobile robotics and autonomous systems.

VIII. FUTURE WORK

Training the model on original LiDAR data with point-wise labels We attempted to simulate LiDAR-like point clouds by processing the egobody dataset. While this approach showed promising results, a better alternative would be to train the model on the original LiDAR point cloud that has per-point annotations for performing human semantic or instance segmentation.

Finetuning Mask-3D on a combination of dense and sparse data We observed a degradation in the model's ability to segment humans in dense point clouds after we fine-tuned the model on sparser point clouds, since the fine-tuning dataset was skewed, having only sparse point clouds. A potential method of avoiding this is to create a new dataset for fine-tuning that is balanced in terms of the sparse and dense point clouds.

Finetuning Human-3D on sparse point clouds While our project aims to solve the problem of performing semantic segmentation of humans from sparse point clouds, future work can involve solving Multi-Human Body-Part Segmentation in sparse point clouds. One possible approach is to fine-tune Human-3D on sparse point clouds.

Introducing rotation invariance to Mask-3D and Human-3D Through our ablation studies, we discovered that Mask3D and Human3D require humans in the point clouds to be aligned in a manner similar to the egobody dataset. Future efforts to make these models rotation invariant are worth exploring through approaches like augmenting the training dataset with a larger range of rotations, or improving the model architectures to better predict the rotation of point clouds in the input example and align it to a suitable canonical axis.

ACKNOWLEDGMENT

We express our profound appreciation to the Robotics Institute at Carnegie Mellon University for providing the necessary resources and environment for this research. We are particularly grateful to Prof. Shubhum Tulsiani and Prof. Ioannis Gkioulekas whose expert guidance was essential during this project's development. Additionally, we thank our teaching assistants for their dedicated support and valuable feedback, which were critical to our success. Their efforts were key in addressing the challenges of implementing this work.