IEEE/SICE International Symposium on System Integration (SII), Munich, Germany, 2025.

Person Segmentation and Action Classification for Multi-Channel Hemisphere Field of View LiDAR Sensors

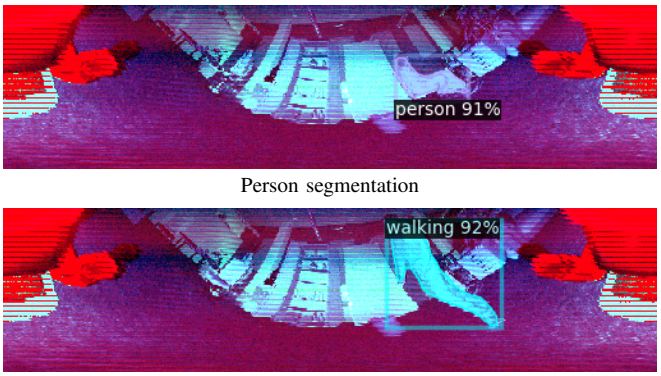
Svetlana Seliunina*, Artem Otelepko*, Raphael Memmesheimer, and Sven Behnke

***Abstract*— Robots need to perceive persons in their surroundings for safety and to interact with them. In this paper, we present a person segmentation and action classification approach that operates on 3D scans of hemisphere field of view LiDAR sensors. We recorded a data set with an Ouster OSDome-64 sensor consisting of scenes where persons perform three different actions and annotated it. We propose a method based on a MaskDINO model to detect and segment persons and to recognize their actions from combined spherical projected multi-channel representations of the LiDAR data with an additional positional encoding. Our approach demonstrates good performance for the person segmentation task and further performs well for the estimation of the person action states walking, waving, and sitting. An ablation study provides insights about the individual channel contributions for the person segmentation task. The trained models, code and dataset are made publicly available.**

I. Introduction

Perceiving persons in the environment is a crucial task for many applications, such as autonomous driving, service robots, and smart buildings. Light Detection and Ranging (LiDAR) sensors are promising for the detection and segmentation of persons in the surrounding for various reasons: i) LiDAR measurements are more reliable than camera-based depth estimates. Hence, they allow for robust detection and precise localization. ii) As LiDAR sensors are actively transmitting laser beams and interpreting their reflections, they remain largely unaffected by changes in lighting conditions and function in complete darkness. iii) LiDAR sensors don't capture direct personal information like facial details, which can improve acceptance by the general public.

Person detection and segmentation are well studied for various sensor data modalities like RGB images [1]–[3], RGB-D images [4], [5], Infrared (IR) images [6]–[8], thermal images [7], [9], [10] and 2D-LiDAR sensors [11]–[13]. With the creation of larger datasets for semantic segmentation of 3D-LiDAR sensors [14] methods for the semantic segmentation of relevant classes such as cars and persons became of increasing interest [15]–[17]. A point-level segmentation allows for precise person localization and further serves as input for tracking [18] and human pose estimation [19]. Especially in the context of assistive service robots, the understanding of the activities of the surrounding persons is of interest. While current LiDAR-based approaches focus solely on the detection and segmentation, we aim to extend the understanding to an activity level.



Person segmentation

Action classification

Fig. 1. We segment persons and classify their actions from spherical projected 2D representations of multi-channel hemisphere FoV LiDAR.

Recent research further has demonstrated that models trained in a supervised fashion could potentially leak private information from training data [20], [21]. We argue that when developing approaches based on sensors that do not measure high-resolution images in the first place, it is impossible to leak such private information.

Hemisphere LiDAR sensors like the Ouster OSDome cover 180° FoV with 64 or 128 laser beams with range, signal, reflectivity and Near Infrared (NIR) channels. As shown in Fig. 1, its measurements can be represented in sensor coordinates as 2D matrix with rotation angle and beam number as coordinate axes. The measurement directions are not evenly distributed across the hemispheric FoV, but are closer together near the axis of rotation and further apart at the periphery, which may pose a challenge for person detection.

In this paper, we explore if recent 2D image detection and segmentation models are capable of dealing with these unevenly distributed beams for person segmentation and action classification tasks.

The contributions of the paper are as follows:

- We provide an annotated dataset for person segmentation and action classification acquired from a hemisphere field of view LiDAR sensor.
- We present a person segmentation approach operating on the combined channel representations of the LiDAR data and further extend this approach to estimate action states of a person, such as walking, waving, and sitting.
- The model, code and dataset are made publicly available

*Equal contribution

This work was funded by grant BE 2556/16-2 (Research Unit FOR 2535 Anticipating Human Behavior) of the German Research Foundation (DFG)

All authors are with the Autonomous Intelligent Systems group, Computer Science Institute VI – Intelligent Systems and Robotics, Lamarr Institute for Machine Learning and Artificial Intelligence, and Center for Robotics, University of Bonn; memmesheimer@ais.uni-bonn.de

to the community on Github[1].

II. Related Work

Detecting persons from sensor data streams is a well-established research topic. In the following, we review the state of the art in person detection from various sensor data modalities and put special emphasis on privacy-preserving person detection methods. Person detection and segmentation methods have a strong focus on image-based methods. Historically, persons were detected with Haar feature cascade detectors [22] and histogram of gradient [23] methods. With increasing classification performance of 2D-Convolutional Neural Network (CNN) they have soon been extended to detection [1], [2] and segmentation models [14], [24].

Sensors for privacy-preserving person detection range from LiDAR to thermal cameras to specially developed sensors. The advances from image-based person detection have always been adapted to other sensors like 2D-LiDAR [11] and 3D-LiDAR [25]. Günter et al. [26] present a privacy-preserving person detection approach for solid state LiDAR sensors. Dubail et al. [6] proposed privacy-preserving person detection using ultra-low resolution IR cameras. The ChaLearn Looking At People Challenge [7] focuses on depth gathered from an RGB-D camera and thermal images for the benchmarking of identity preserving approaches. The annotations are on a bounding box level, whereas we provide annotations on a pixel level. In the related challenge, the best performing approaches utilized the thermal images with a combination of Faster R-CNN [1] + Faster R-DCN (ResNet-50) [27] and a soft Non-maximum Suppression (NMS).

In many instances, especially for Human Robot Interaction (HRI), one might be interested not only in the persons' location but also in their gestures or activities. Droeschel *et al.* [28] track persons of interest with a 2D LiDAR and utilize a Time of Flight (ToF) camera to extract pointing gestures from a mobile manipulating platform. Their approach, due to the usage of ToF is not usable to preserve privacy, as the sensor leaks facial attributes. For gesture recognition, they segment the body into multiple parts and then estimate the face centroid, elbow position and hand position. Vectors between the body keypoints are utilized to estimate a showing and pointing gesture.

For recognizing activities directly from videos, spatio-temporal models have been proposed. SlowFast [29] proposes a two stream approach, one stream (high frame rate) focuses on extracting temporal features and the second stream (low frame rate) focuses on the extraction of detailed spatial semantics. Multi-modal approaches that generalize well across different sensor modalities have been proposed for the action recognition task in a supervised setting [30] and for one-shot inference following a semi-supervised setting [31]. In contrast to the above-mentioned approaches, which focus on sequence classification, our proposed approach can also localize the activities.

[1]https://github.com/AIS-Bonn/lidar_person_action_detection

Fig. 2. The dataset collection robot setup.

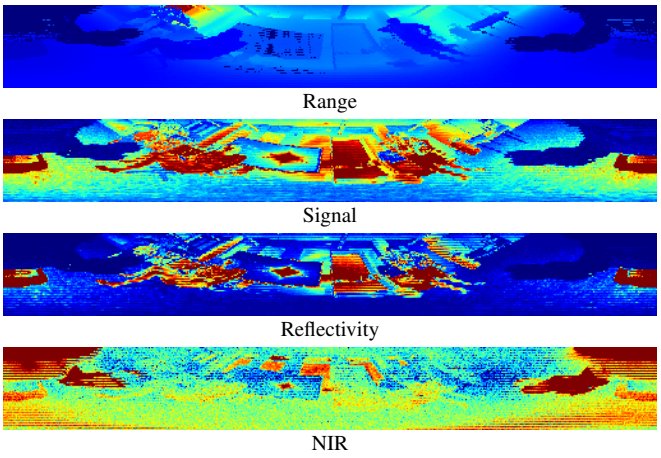


Fig. 3. Example data from the individual Ouster OSDome-64 channels.

3D LiDAR semantic segmentation datasets such as SemanticKITTI [14] often focuses on automotive scenes. In contrast, our focus is on indoor person segmentation and additional action classification.

III. Dataset

The dataset is collected with an Ouster OSDome-64 LiDAR attached on a TIAGo++ omnidirectional mobile robot. The LiDAR sensor is attached in the front and tilted slightly downwards to capture the environment in front of the robot, as depicted in Fig. 2.

The Ouster OSDome-64 LiDAR provides four channels, shown in Fig. 3: 1) the distance of the measured surface in mm, 2) the signal intensity (number of photons in the signal return measurement) 3) reflectivity (scaled intensity based on measured range and sensor sensitivity at that range) 4) NIR (photons related to natural environmental illumination).

The dataset contains 442 scans of persons in different action states, such as walking, waving, and sitting. The dataset samples are randomly divided once into training, validation and test datasets with proportions of 70/15/15. The dataset consists of scenes of different complexity. In some of them, only one or two persons are present, and

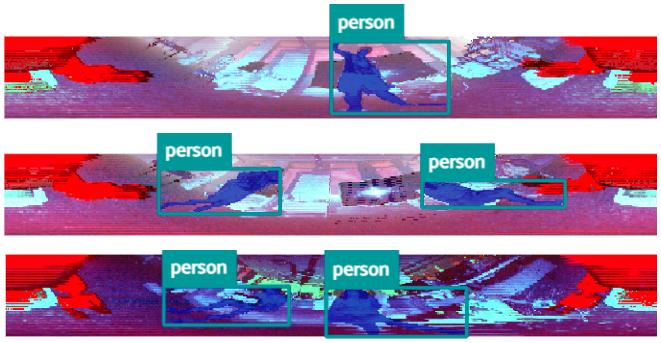


Fig. 4. Person segmentation examples with ground truth masks (blue).

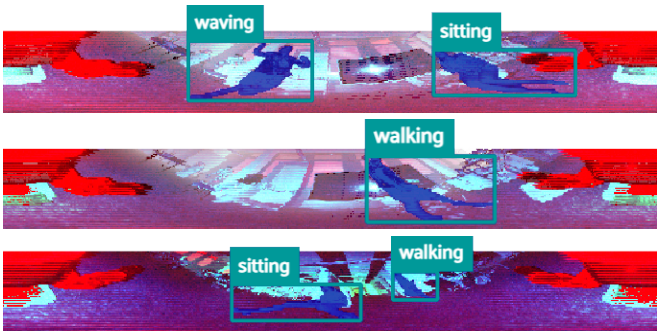


Fig. 5. Action recognition examples with ground truth mask (blue).

they are clearly visible. In others, more people are present in the background. Persons of different body types and sizes with different clothes are present in the data. The background itself changes from scene to scene, and people are often partially occluded by various objects. It is also worth mentioning that the data is collected with both the stationary and the moving robot.

During the walking action, people walked in different directions but were instructed not to raise their hands. While waving, the subjects could stand or walk, but one or both hands were always raised above waist height and moving. Finally, for the sitting action, people could sit in different positions, be occluded by a table and move the chair. In each action sequence, sensor data was collected ten times.

To support the labeling, we used a semi-automatic labeling approach based on SegmentAnything [32] and manually corrected the masks. Examples for person segmentation and action classification are shown in Figs. 4 and 5, respectively.

IV. Method

We adapted a MaskDINO model [33] to jointly train a person detection and segmentation model on our proposed multi-channel representation.

A. Representation

Each measurement channel has advantages and disadvantages in terms of perceptual separation of individuals, depending on the situation. None of the channels consistently provides universal perceptual separation for individuals, though. Hence, we incorporated all LiDAR channels in the person perception, thereby improving the separation

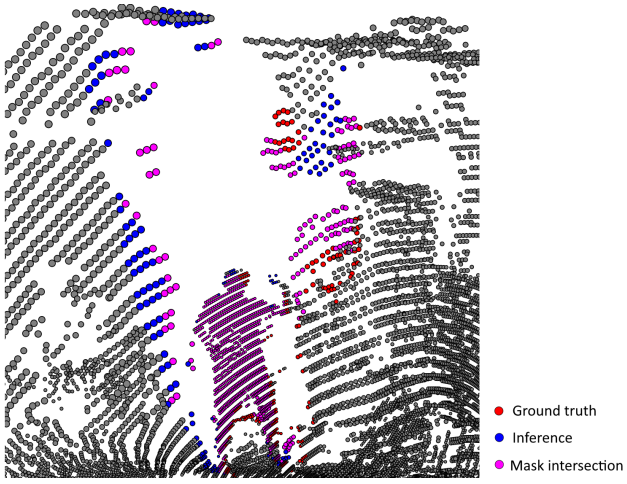


Fig. 6. Point cloud with ground truth and segmentation masks.

of individuals, as the quality of the masks was significantly influenced by this.

After conducting several experiments with various combinations of channels, we decided to utilize reversed range data as the opacity channel for visualization. This technique enabled us to render objects that are located far from the sensor transparent in tools that support four channel images, and to achieve subjectively good perceptual separation of the individuals in the image in tools that do not correctly support four channel images. This was advantageous mainly for simplifying the process of manual pixel-level annotation.

The resulting image representation consists of NIR, reflectivity, signal and the reversed range in separate channels resulting in a four channel image of size 512×64.

From the range image and LiDAR metadata, we reconstructed the Euclidean coordinates of each corresponding pixel. The sample image together with the ground truth and inferred masks are presented in Fig. 6. Errors in the ground truth mask are the consequences of using the semiautomatic approach for labeling on the four-channel image.

Positional encoding of the XYZ point coordinates were added as additional input channels, resulting in a seven-channel input image. The positional encoding is computed using Ouster Sensor SDK as follows:

$$
\begin{aligned}
x &= (r - |n|)\cos(\theta_{enc} + \theta_{azi})\cos(\phi) + x_n \cos(\theta_{enc}),\\
y &= (r - |n|)\sin(\theta_{enc} + \theta_{azi})\sin(\phi) + x_n \cos(\theta_{enc}),\\
z &= (r - |n|)\sin(\phi) + z_n,
\end{aligned}
$$

with

$$
|n| = \sqrt{x_n^2 + z_n^2}, \quad \theta_{enc} = 2\pi\left(1 - \frac{i}{w}\right),
$$

$$
\theta_{azi} = -2\pi\frac{\alpha_i}{360}, \quad \phi = 2\pi\frac{\beta_i}{360},
$$

where r is the range value of the measurement ID i. Parameters x_n, z_n describe the distance from the center of the LiDAR origin coordinate frame to its front optics. w denotes the scan width. α_i, β_i denote the azimuth angle and altitude angle of beam i, respectively.

TABLE I

TRAINING RESULTS FOR PERSON DETECTION

Pos. enc.*	Precision	F1-score	Precision	F1-score
	Frozen Backbone		Backbone also trained	
✗	0.98	0.93	0.98	0.97
✓	0.98	0.95	0.98	0.97

*positional encoding

B. Model

We adapted MaskDINO [33] with a SwinL transformer [34] backbone. The model operates on four-channel input images, or seven channels if the positional encoding is used, by employing a convolutional encoding layer before the backbone in both the trainer and the predictor. The full representations are then fed to the model during training and inference. The same model is applicable for person segmentation and action detection. The hyperparameters in our approach were based on the MaskDINO configuration with additional changes to incorporate the images with required number of channels and resizing to 512×64 pixels. Horizontal random flip was used as data augmentation. The models were trained on a Nvidia RTX 3090 GPU, which resulted in approximately 35 minutes training time for 5,000 iterations.

V. EVALUATION

We evaluate our approach for person segmentation and action classification on the proposed dataset. We further give an ablation study on the channel contributions and the effect of the positional encoding for the person segmentation task. Finally, we analyze the applicability on an actual robot setup for online person segmentation and action classification.

A. Person Segmentation

The first set of experiments performs person segmentation with all four channels from the constructed dataset. Our training procedure initializes MaskDINO with SwinL transformer backbone with weights pretrained on COCO 2017.

The models were trained several times for 5,000 iterations with a step at 4,000 iterations, which multiplied the learning rate by 0.1. We trained the models with base learning rates 1×10^{-3}, 1×10^{-4}, and 1×10^{-5} and found that models trained with a 1×10^{-4} learning rate provided the most promising and consistent results. After determining the base learning rate, we conducted several experiments with a frozen and unfrozen backbone.

Using a frozen backbone provides good results, but the F1-score varies between 0.6 and 0.9. To further improve the results, we decided to unfreeze the backbone. A backbone multiplier of 1×10^{-5} provided the best results with an F1-score higher than 0.9.

The implementation of positional encoding allowed us to improve the performance of the trained models, resulting in higher precision and F1-score.

The best attempts to train the model for person segmentation with and without positional encoding are listed in Table I. These results correspond to the final iteration of training for 5,000 iterations with the base learning rate of 1×10^{-4}.

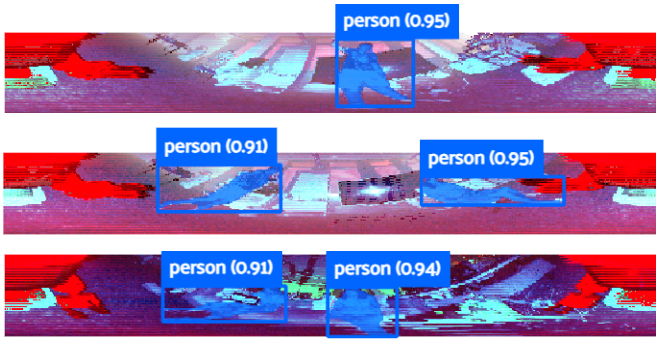


Fig. 7. Person detection examples with inferred mask (blue).

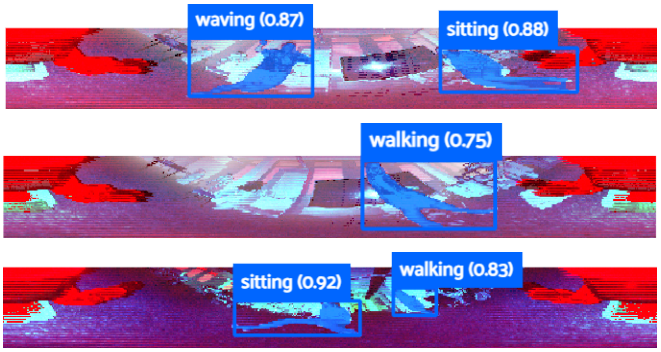


Fig. 8. Action detection examples inferred mask (blue).

TABLE II

ACTION DETECTION INITIALIZED WITH PERSON DETECTION WEIGHTS.

Class	Pos. enc.*	Precision	F1-score	Precision	F1-score
		Frozen Backbone		Non Frozen Backbone	
sitting	✗	0.97	0.95	1.00	0.97
walking	✗	0.88	0.87	0.85	0.85
waving	✗	0.83	0.89	0.86	0.91
w. avg**		0.91	0.91	0.92	0.92
sitting	✓	0.95	0.95	0.95	0.96
walking	✓	0.89	0.91	0.93	0.95
waving	✓	0.92	0.92	0.86	0.91
w. avg**		0.93	0.93	0.92	0.94

*positional encoding, **weighted average

The models trained with the base learning rate equal to 1×10^{-4} and unfrozen backbone with the backbone multiplier equal to 1×10^{-5} provided good results and were used in the following experiments. Fig. 7 depicts person segmentation on the sample images from Fig. 4 using the model without positional encoding.

B. Action Classification

For domestic service robot applications, the action classification of the surrounding persons might be of interest. We initialized the weights from our person segmentation models and fine-tuned them to the action classification task.

Similarly to the person segmentation experiment, the models were trained several times for 5,000 iterations with a step

TABLE III

ACTION DETECTION MODELS INITIALIZED WITH MASKDINO WEIGHTS

Class	Pos. enc.*	Precision	F1-score	Precision	F1-score
		Frozen Backbone		Non Frozen Backbone	
sitting	✗	0.97	0.94	1.00	0.82
walking	✗	0.96	0.94	1.00	0.77
waving	✗	0.89	0.91	0.73	0.73
w. avg**		0.95	0.93	0.92	0.78
sitting	✓	0.93	0.95	1.00	0.97
walking	✓	1.00	0.94	0.88	0.82
waving	✓	0.83	0.87	0.93	0.95
w. avg**		0.92	0.93	0.94	0.92

*positional encoding, **weighted average

at 4,000 iterations, which multiplied the learning rate by 0.1.

The base learning rate in the experiments was chosen to be 1×10^{-4} and 1×10^{-5}. The model with learning rate 1×10^{-5} performed slightly worse with an F1-score lower than 0.9, so we set the learning rate to 1×10^{-4}. After determining the base learning rate, we conducted several experiments with a frozen and unfrozen backbone. The backbone learning rate multiplier for the unfrozen model was set to 1×10^{-5}.

The best attempts to train the model for action classification are listed in the Table II. These results correspond to the final iteration of training for 5,000 iterations with the base learning rate of 1×10^{-4}. Frozen and unfrozen backbones performed equally in this experiment. The positional encoding had a positive influence. Fig. 8 depicts action classification on the sample images from Fig. 5 using the best model without positional encoding.

Another approach was to initialize the training with weights of a pre-trained on the COCO 2017 dataset MaskDINO model with SwinL transformer backbone instead of using the weights of a trained model for the person detection task. To compare the results of different approaches fairly, we trained the models for 10,000 iterations with a base learning rate 1×10^{-4} and a step at 8,000 iterations, which multiplied the learning rate by 0.1. The best attempts are shown in Table III.

Contrary to training the model using the weights of a trained model for the person detection task, this experiment provided the best results with frozen backbone. The results of this experiment, however, were significantly less consistent than the previous one.

C. Channel Contribution Ablation

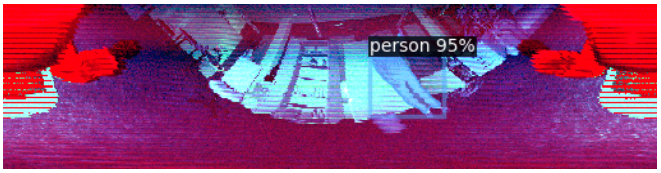
To understand the channel contributions of the proposed representation, we trained the models with a random fixed seed excluding different image channels and measured the performance of the model at the end of the training. We conducted the ablation on the person segmentation task by utilizing the hyperparameters to match the best experiment from Table I. The results for each excluded channel are shown in Table IV. NIR and signal channels have only a minor contribution while reflectivity and range channels data have higher contribution. Our ablation study indicates that

TABLE IV

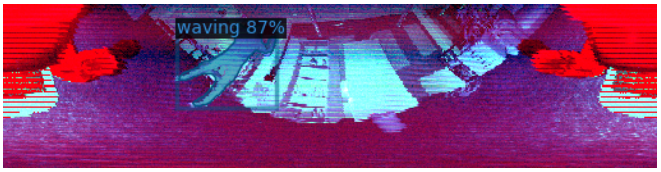
ABLATION ON CHANNEL CONTRIBUTION (PERSON DETECTION TASK)

Excluded channel	Pos. enc.*	Result		
		Precision	Recall	F1-score
-	✗	0.89	0.42	0.57
NIR	✗	0.97	0.94	0.95
Reflectivity	✗	0.97	0.74	0.84
Signal	✗	0.96	0.78	0.86
Range	✗	0.65	0.67	0.66
-	✓	0.83	0.63	0.72
NIR	✓	0.94	0.85	0.89
Reflectivity	✓	0.60	0.31	0.41
Signal	✓	0.93	0.85	0.89
Range	✓	0.58	0.33	0.42

*positional encoding



Person segmentation



Action classification

Fig. 9. Online detection experiment on the TIAGO++ Omni robot platform.

the contribution to the performance of the model of different channels differs between the cases with positional encoding and without it. The ablation study shows that the model with fewer channels produces a better result than the one with all of them for both models. One possible explanation is that the one-layer convolutional encoder does not deal well with transforming our multi-channel input to the desired MaskDINO input format. In future work, we may explore different encoders and their contribution.

D. Online Application on Robot

To assess our methods for real-time application, we tested the inference of the best models for person segmentation (Table I) and action classification (Table II) on the data received from an Ouster OSDome-128 LiDAR which yields images of size 512×128. Note, the LiDAR sensor for online experiments is slightly different and has the double amount of laser beams, demonstrating that our proposed approach generalizes across different versions of the sensor without adaptations. The average inference rate on a Zotac ZBOX QTG7A4500 equipped with an Nvidia RTX A4500 16GB GPU yields 11 Hz with and without positional encoding, which is suitable e.g. as input for person tracking approaches. Examples of the results for person segmentation and action classification are shown in Fig. 9. The results indicate that our approach generalizes well for images of different sizes

from different sensors and is suitable for online application.

VI. Conclusion

We presented a person segmentation and action classification approach for multi-channel data of hemisphere FoV 3D LiDAR sensors. A dataset with segmentation-level annotations on multi-channel LiDAR measurements has been collected and annotated and a model based on MaskDINO has been adapted and trained to estimate person segments and has further been shown to be capable of estimating three different action classes relevant for HRI. An ablation study provided insights into the channel contributions for the person segmentation model, demonstrating that the range and reflectivity channels as well as the positional encoding contribute significantly to the performance. Our approach demonstrated good performance for both, the person segmentation and the classification of three different person states. It is real-time capable and applicable to a sensor with more LiDAR beams.

References

[1] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] S. Shin, J. Kim, E. Halilaj, and M. J. Black, "WHAM: Reconstructing world-grounded humans with accurate 3D motion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2070–2080.

[4] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5636–5643.

[5] Z. Xu, X. Zhan, Y. Xiu, C. Suzuki, and K. Shimada, "Onboard dynamic-object detection and tracking for autonomous robot navigation with RGB-D camera," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 1, pp. 651–658, 2023.

[6] T. Dubail, F. A. Guerrero Peña, H. R. Medeiros, M. Aminbeidokhti, E. Granger, and M. Pedersoli, "Privacy-preserving person detection using low-resolution infrared cameras," in *European Conference on Computer Vision (ECCV) Workshops*, ser. Lecture Notes in Computer Science, vol. 13805, Springer, 2022, pp. 689–702.

[7] A. Clapés, J. C. S. J. Júnior, C. Morral, and S. Escalera, "ChaLearn LAP 2020 challenge on identity-preserved human detection: Dataset and results," in *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2020, pp. 801–808.

[8] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 45, no. 8, pp. 10 535–10 554, 2023.

[9] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *24th European Symposium on Artificial Neural Networks (ESANN)*, 2016.

[10] C. Tian, Z. Zhou, Y. Huang, G. Li, and Z. He, "Cross-modality proposal-guided feature mining for unregistered RGB-thermal pedestrian detection," *IEEE Transactions on Multimedia (TMM)*, 2024.

[11] L. Beyer, A. Hermans, T. Linder, K. O. Arras, and B. Leibe, "Deep person detection in two-dimensional range data," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 3, pp. 2726–2733, 2018.

[12] D. Jia, A. Hermans, and B. Leibe, "DR-SPAAM: A spatial-attention and auto-regressive model for person detection in 2D range data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 270–10 277.

[13] H. Yang, Y. Yang, C. Yao, C. Liu, and Q. Chen, "Li2Former: Omni-dimension aggregation transformer for person detection in 2-D range data," *IEEE Tr. on Instrumentation and Measurement (TIM)*, 2024.

[14] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of LiDAR sequences," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9296–9306.

[15] Z. Yan, T. Duckett, and N. Bellotto, "Online learning for 3D LiDAR-based human detection: Experimental analysis of point cloud clustering and classification methods," *Autonomous Robots*, vol. 44, no. 2, pp. 147–164, 2020.

[16] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

[17] F. Hong, L. Kong, H. Zhou, X. Zhu, H. Li, and Z. Liu, "Unified 3D and 4D panoptic segmentation via dynamic shifting networks," *IEEE Tr. on Pattern Analysis and Machine Intelligence (PAMI)*, 2024.

[18] H. Wang, B. Wang, B. Liu, X. Meng, and G. Yang, "Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle," *Robotics and Autonomous Systems (RAS)*, vol. 88, pp. 71–78, 2017.

[19] J. Shotton, T. Sharp, A. Kipman, A. W. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[20] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the GAN: information leakage from collaborative deep learning," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017, pp. 603–618.

[21] H. Li, D. Guo, W. Fan, M. Xu, and Y. Song, "Multi-step jailbreaking privacy attacks on ChatGPT," *preprint arXiv:2304.05197*, 2023.

[22] P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 511–518.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[24] T. Cortinhal, G. Tzelepis, and E. Erdal Aksoy, "SalsaNext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds," in *15th International Symposium on Advances in Visual Computing (ISVC)*, 2020, pp. 207–222.

[25] K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura, "Pedestrian recognition using high-definition LIDAR," in *IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 405–410.

[26] A. Günter, S. Böker, M. König, and M. Hoffmann, "Privacy-preserving people detection enabled by solid state lidar," in *16th Int. Conference on Intelligent Environments (IE)*, IEEE, 2020.

[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[28] D. Droeschel, J. Stückler, D. Holz, and S. Behnke, "Towards joint attention for a domestic service robot - person awareness and gesture recognition using time-of-flight cameras," in *IEEE Int. Conference on Robotics and Automation (ICRA)*, 2011, pp. 1205–1210.

[29] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.

[30] R. Memmesheimer, N. Theisen, and D. Paulus, "Gimme Signals: Discriminative signal encoding for multimodal activity recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 394–10 401.

[31] R. Memmesheimer, N. Theisen, and D. Paulus, "SL-DML: Signal level deep metric learning for multimodal one-shot action recognition," in *25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2020, pp. 4573–4580.

[32] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[33] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H. Shum, "Mask DINO: towards A unified transformer-based framework for object detection and segmentation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3041–3050.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.