Human-centric Scene Understanding for 3D Large-scale Scenarios

Yiteng Xu[1,*], Peishan Cong[1,*], Yichen Yao[1,*],
Runnan Chen[2], Yuenan Hou[3], Xinge Zhu[4], Xuming He[1], Jingyi Yu[1], Yuexin Ma[1,†]
[1] ShanghaiTech University [2] The University of Hong Kong
[3] Shanghai AI Laboratory [4] The Chinese University of Hong Kong
{xuyt1,congpsh,yaoych,mayuexin}@shanghaitech.edu.cn

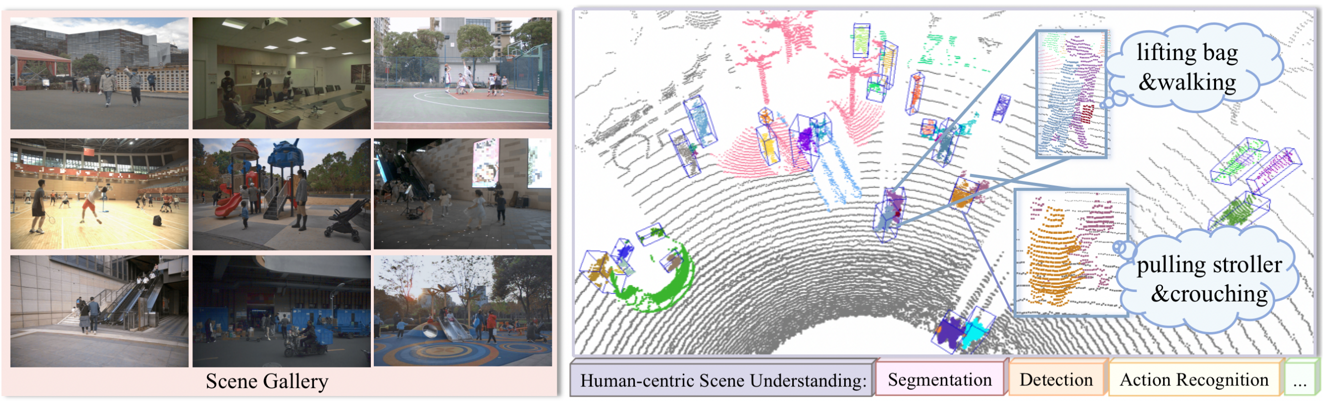

Figure 1. The left shows several scenes captured in HuCenLife, which covers diverse human-centric daily-life scenarios. The right demonstrates rich annotations of HuCenLife, which can benefit many tasks for 3D scene understanding .

Abstract

Human-centric scene understanding is significant for real-world applications, but it is extremely challenging due to the existence of diverse human poses and actions, complex human-environment interactions, severe occlusions in crowds, etc. In this paper, we present a large-scale multi-modal dataset for human-centric scene understanding, dubbed HuCenLife, which is collected in diverse daily-life scenarios with rich and fine-grained annotations. Our HuCenLife can benefit many 3D perception tasks, such as segmentation, detection, action recognition, etc., and we also provide benchmarks for these tasks to facilitate related research. In addition, we design novel modules for LiDAR-based segmentation and action recognition, which are more applicable for large-scale human-centric scenarios and achieve state-of-the-art performance. The dataset and code can be found at https://github.com/4DVLab/HuCenLife.git.

1. Introduction

Human-centric scene understanding in 3D large-scale scenarios is attracting increasing attention [13, 11, 42, 31], which plays an indispensable role in human-centric applications, including assistive robotics, autonomous driving, surveillance, human-robot cooperation, *etc*. It is often confronted with substantial difficulties since these human-centric scenarios usually have the attributes of various subjects with different poses, fine-grained human-object interactions, and challenging localization and recognition with occlusions. Moreover, current state-of-the-art perception methods heavily rely on large-scale datasets to achieve good performance. Therefore, to promote the research of human-centric scene understanding, the collection of large-scale datasets with rich and fine-grained annotations is required urgently, which is difficult but of great significance.

In previous work, many studies target on the scene un-

*Equal contribution. † Corresponding author. This work was supported by NSFC (No.62206173), Natural Science Foundation of Shanghai (No.22dz1201900), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), Shanghai Engineering Research Center of Intelligent Vision and Imaging.

derstanding based on the input of image or video [2, 37, 17, 59], which are not applicable to real-world applications due to the limited 2D visual representations. Afterward, some works pay attention to the static indoor-scene understanding [12, 1, 5] based on the pre-scanned RGB-D data, which are not suitable for the research of real-time perception. Recently, more and more outdoor multi-modal datasets [6, 49] are released equipped with LiDAR point clouds. They provide detailed annotations under complex outdoor scenes, while they often focus on the vehicle-dominated traffic environment and neglect the more challenging human-centric daily-life scenarios. Although the dataset STCrowd [11] appears lately, it focuses on the detection task of dense pedestrian scenes, lacking varied human activities and diversified annotations. Consequently, the dataset with rich and fine-grained annotations for human-centric understanding in long-range 3D space is crucial and insufficient.

In this paper, to facilitate the research of human-centric 3D scene understanding, we collect a large-scale multi-modal dataset, namely HuCenLife, by using calibrated and synchronized camera and LiDAR. Specifically, the dataset captures 32 multi-person involved daily-life scenes with rich human activities and human-object interactions. Various indoor and outdoor scenarios are both included. For the annotation, we provide fine-grained labels including instance segmentation, 3D bounding box, action categories, and continuous instance IDs, which can benefit various 3D perception tasks, such as point cloud segmentation, detection, action recognition, Human-Object Interaction (HOI) detection, tracking, motion prediction, *etc*. In this paper, we provide benchmarks for the former three tasks by executing current state-of-the-art methods on HuCenLife and give discussions for other downstream tasks.

In particular, considering the specific characteristics of human-centric scenarios, we propose effective modules to improve the performance for point cloud-based segmentation and action recognition in the complex human-centric environments. First, we model human-human interactions and human-object interactions and leverage their mutual relationships to benefit the classification of points and instances. Second, to solve the problem of the big scale span of objects in daily-life scenarios, we exploit multi-resolution feature extraction strategy to aggregate global features and local features hierarchically so that small objects can be better attended. We evaluate our methods and conduct extensive experiments on HuCenLife. Several ablation studies are also conducted to demonstrate the effectiveness of each module and good generalization capability. Our contributions are summarized as follows:

1. We introduce HuCenLife, the first large-scale multi-modal dataset for human-centric 3D scene understanding with rich human-environment interactions and fine-grained annotations.
2. HuCenLife can benefit various human-centric 3D perception tasks, including segmentation, detection, action recognition, HOI, tracking, motion prediction, etc. We provide baselines for three main tasks to facilitate future research.
3. Several novel modules are designed by incorporating fine-grained interactions and capturing features at various resolutions to promote more accurate perception in human-centric scenes.

2. Related Work

2.1. Datasets for 3D Scene Understanding

The RGB-D datasets of indoor scenes dominate the early scene understanding task. ScanNet [12, 1] focuses on object surface reconstruction and semantic segmentation, providing dense and rich annotations for various indoor objects. NTU RGB+D [44] is a human action recognition dataset with corresponding skeleton and action labels. Behave [5] concentrates on human-object interaction with human SMPL models and interactive objects annotations. It can be found that outdoor scenarios are not well explored. Recently, the community has paid attention to traffic scenes for autonomous driving and collect several outdoor multi-modal datasets. KITTI [21], nuScenes [6] and Waymo [50] provide 3D bounding boxes for traffic participants and [6, 3] also offer point-wised semantic segmentation labels. However, these datasets are all vehicle-dominated and neglect human-centric scenarios. STCrowd [11] mainly concentrates on the crowds on campus but lacks the fine-grained segmentation labels and complex human-environment interactions. In order to facilitate the research of human-centric 3D scene understanding, we collect HuCenLife, a multi-modal dataset with various scenarios in human daily life.

2.2. Point Cloud-based Segmentation

Most outdoor point cloud segmentation methods mainly focus on point cloud representations. Point-based methods [39, 40, 67, 52] make the operation on unordered point cloud directly. Voxel-basedmethods [10, 22] utilize efficient sparse convolution to reduce the time complexity. PolarNet [71] and Cylinder3D [76] further consider the non-uniform LiDAR point clouds characteristics and point distribution, and divide the points under the polar coordinate system. [26] adopts the cylinder convolution and proposes a dynamic shifting network for instance prediction. These methods are mainly focusing on automatic driving scenes, while neglecting the counterpart in human-centric scenarios with complex human-object interactions and challenging occlusions.

Another line of segmentation, namely point cloud instance segmentation, also embraces great progress, which can be mainly divided into proposal-based methods and grouping-based methods. Previous proposal-based methods [64, 18, 61] regard the instance segmentation as a top-down pipeline, which first generate proposals and then segment the objects within the proposals. Grouping-based methods [27, 55, 23, 8, 24, 58] adopt the bottom-up strategy. PointGroup [27] aggregates points from original and offset-shifted point sets. DyCo3D [8] and DKNet [58] encode instances into kernels and propose dynamic convolution kernels and then merge the candidates. Considering the imprecise bounding box prediction in proposal-based methods for refinement and the time-consuming aggregation in grouping methods, [43, 48] take each object instance as an instance query and design a query decoder with transformers. However, these methods are applied to structured indoor instances without human involvement and human-environment interactions. Our dataset and proposed method target more on human-human and human-object interactions in large-scale human-centric scenes.

2.3. LiDAR-based 3D Detection

As the mainstream of 3D perception task, 3D detection task has been fully explored, which can be grouped via the point encoding strategies. First, point-based methods [69, 7, 38, 46, 62] extract the geometry information from raw points with sampling and grouping. [53, 56, 4, 51, 30, 19] transform point cloud into range images for detection. Second, voxel-based methods [74, 57, 29, 15, 14, 65, 60] convert raw point clouds to regular volumetric or pillar representations and adopt voxel-based feature encoding. Third, hyper-fusion methods [36, 28, 45, 63, 9, 75] take advantage of both voxels and points and fuse them together to model the hyper encoding. In this paper, we test them on the proposed HuCenLife dataset to provide the benchmark and offer the comprehensive analyses and comparison.

2.4. Action Recognition

Recently, transformer-based methods have dominated the field of action recognition [32, 20]. Many variants based on ViT [17] have been proposed to explore the potential of transformer in video classification, where ViViT [2] extends the two-dimensional patch to three-dimensional tube to model the temporal relation, MTV [59] divides the tube with different time scales to extract the action features with different amplitude of change over time, and TubeViT [37] further samples various sized 3D space-time tubes from the video to generate learnable tokens. However, the common action recognition [47, 44] is annotated in image-level and lacks of instance-level labels, causing these methods hard to be applicable in complex 3D scenarios. In this paper, we introduce point cloud-based instance action recognition task in large-scale scenes and collect the HuCenLife dataset equipped with various instances with different poses and motions, to make the basis for research community.

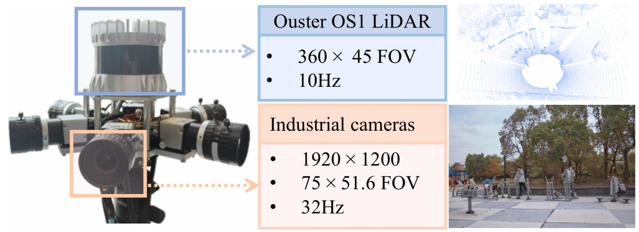

Figure 2. Sensor setup for data collection.

3. HuCenLife Dataset

HuCenLife is the first dataset that emphasizes human-centric 3D scene understanding, containing indoor and outdoor daily-life scenes with rich annotations of human activities, human-human interactions, and human-object interactions, which facilitates the development of intelligent security, assistive robots, human-machine cooperation, *etc*. In this section, we first introduce the data acquisition in Sec.3.1, and then provide important annotation statistics in Sec.3.2, and finally highlight the novelties of HuCenLife by comparing with existing influential datasets in Sec.3.3.

3.1. Data Acquisition

To collect the dataset, we built a Visual-LiDAR Capture System, which mainly consists of one 128-beam Ouster-OS1 LiDAR and six industrial cameras in a circle, as Fig 2 shows. All sensors are tied in fixed positions on the bracket with mechanical synchronization. The LiDAR has a 360° horizon field of view (FOV) ×45° vertical FOV, and each camera has a $75° \times 51.6°$ FOV with 1920×1200 image resolution. For our equipment, LiDAR captures raw point cloud in 10Hz and camera takes pictures in 32Hz.

3.2. Annotation

We manually annotated all humans and these objects with interactions with humans in LiDAR point cloud by referring to the synchronized image. We select one frame per second for labeling and finally obtain 6,185 frames (103 minutes) of annotated LiDAR point cloud. For each target, we provide four kinds of annotations, *i.e.*, point cloud-based instance segmentation, 3D bounding box, human action classification, and tracking ID across consecutive frames, like Fig 1 shows. In HuCenLife, there are 65,265 human instances in total, including 58,354 adults and 6,911 children, and 31,303 human-interacted objects. There are 20 categories of objects and 12 kinds of human actions. Specifically, the HuCenLife dataset is collected in 15 distinguished locations with 32 human-centric daily-life scenes, including playground, shopping mall, campus, park, gym, meeting room, express station, *etc*. For each scene, there are 11 persons on average with multiple interacted objects, and for some complex scenes, there are about 70 persons. The diverse density distributions in HuCenLife bring

Table 1. Comparison with related datasets for 3D scene understanding. There are some abbreviations, where “pc” denotes LiDAR point cloud, “ins. seg.” means instance segmentation, “bbx” is bounding box, and “inter. obj.” denotes objects having interactions with humans.

Dataset	Data Modality	LiDAR Beam	Point Cloud Frame	Person Number	Person Per Frame	Scenes		Annotation Content			Annotation Targets	
						indoor	outdoor	ins. seg.	3D bbx	action	multi-person	inter. obj.
ScanNet[12]	RGBD	-	-	-	-	✓	✗	✓	✓	✗	✗	✗
S3DIS[1]	RGBD	-	-	-	-	✓	✗	✓	✓	✗	✗	✗
SUN RGB-D[47]	RGBD	-	-	-	-	✓	✗	✗	✓	✗	✗	✗
NTU RGB+D[44]	RGBD	-	-	-	-	✓	✗	✗	✗	✓	✗	✗
BEHAVE[5]	RGBD	-	15.8k	15.8k	1	✓	✗	✓	✗	✓	✗	✓
SemanticKITTI[3]	pc	64	43k	9.7k	0.2	✗	✓	✓	✗	✗	✓	✗
KITTI[21]	image&pc	64	15k	4.5k	0.3	✗	✓	✓	✓	✗	✓	✗
Waymo[50]	image&pc	64	230k	2.8M	12	✗	✓	✓	✓	✗	✓	✗
nuScenes[6]	image&pc	32	40k	208k	5	✗	✓	✓	✓	✗	✓	✗
STCrowd[11]	image&pc	128	11k	219k	20	✗	✓	✗	✓	✗	✓	✗
HuCenLife	image&pc	128	6.1k	65k	11	✓	✓	✓	✓	✓	✓	✓

challenges for related research. More detailed annotation introductions are in the supplementary material.

3.3. Characteristics

We introduce the basic information of HuCenLife and compare it with related popular datasets in Table 1. In particular, we conclude four highlights of our dataset below.

Large-scale Dynamic Scenarios. Benefiting from the long-range-sensing and light-independent properties of LiDAR, HuCenLife contains data of diverse large-scale scenes day and night. Unlike indoor datasets [12] where the scene is pre-scanned and has only static objects, HuCenLife provides online captured multi-modal visual data of dynamically changing scenes with dynamic people, objects, and background. Furthermore, the density of humans and objects is changing from a few to dozens in distinct scenes. The visual data in such diverse dynamic scenarios has huge significance for developing mobile robots.

Abundant Human Poses. Different from current traffic or crowd datasets [50, 6, 11], where people only act as pedestrians walking or standing on the road, HuCenLife pays attention to daily-life scenarios, where people have rich actions, such as doing exercise, crouching down, dancing, running, riding, *etc*. In particular, HuCenLife contains thousands of children samples, which are never concerned in previous datasets. Such complex scenarios with high-degree freedom of human poses bring challenges for accurate perception and recognition.

Diverse Human-centric Interactions. Apart from abundant self-actions of humans, HuCenLife also includes rich human-human interactions (hugging, holding hands, holding a baby, *etc*.) and human-object interactions (riding a bike, opening the door, carrying a box, *etc*.). What's more, there are some extremely complex human-human-object interactions, such as playing basketball, having a meeting in a room, *etc*., which require the participation of multiple persons and objects. HuCenLife is unique for containing diversified interaction data in a variety of scenes, which is significant for the research of human-machine cooperation and boosts the development of service robots.

Rich Annotations. HuCenLife provides rich fine-grained annotations, which can benefit many perception tasks, such as point cloud segmentation, 3D detection, 3D tracking, action recognition, HOI, motion prediction, *etc*. In particular, due to complex scene contents, the annotation process of HuCenLife is much more difficult than others. A well-trained annotator usually spends 25min on average for labeling one frame of LiDAR point cloud in our dataset.

3.4. Privacy Preservation

We strictly obey the privacy-preserving rules. We mask all sensitive information, such as the faces of humans and locations, in RGB images. LiDAR point clouds without any texture and facial information naturally protect the privacy.

4. Various Downstream Tasks

As mentioned above, our dataset can benefit numerous human-centric 3D perception tasks. We conduct three main tasks on HuCenLife based on the LiDAR point cloud, including human-centric instance segmentation, human-centric 3D detection, and human-centric action recognition, and provide the baseline methods. Particularly, novel methods are proposed for instance segmentation and action recognition, respectively, to tackle the difficulties of large-scale human-centric scenarios. In what follows, we present details of these tasks with extensive experiments in order.

5. Human-centric Instance Segmentation

For LiDAR point cloud-based semantic instance segmentation, the input is expressed as $P \in \mathcal{R}^{N\times4}$, which involves N points with the 3D location and reflection intensity (x, y, z, r). The task is to assign each point to a category and then output a set of object instances with their corresponding semantic labels.

5.1. Method

For human-centric scenes, people have diverse pose types and may stay together with occlusions. Moreover,

Table 2. Instance segmentation results on HuCenLife dataset.

	person	motorbike	table	box	cart	seesaw	basketball	fitness equ	cabinet	baby	blackboard	staircase	slide	scooter	computer	backpack	obj in hand	chair	spring car	ground	mIOU	AP50	AP25
Voxel-DSNet [26]	66.9	16.1	20.7	22.6	16.3	12.6	5.4	11.9	1.1	25.7	58.3	8.5	72.7	33.3	24.6	20.7	1.6	8.6	3.1	97.9	26.4	2.6	7.1
Cylinder-DSNet [26]	72.3	12.9	23.8	28.6	18.9	25.2	5.8	4.7	6.8	23.4	90.2	21.4	67.9	37.2	15.5	23.9	3.5	14.1	2.8	97.9	29.8	1.2	7.6
DKNet [58]	75.6	52.7	5.3	26.3	35.8	65.6	0.0	14.6	0.6	39.7	93.9	0.0	95.1	48.5	13.1	9.8	14.6	8.1	3.4	98.0	35.0	11.1	14.0
SoftGroup [55]	80.0	32.6	4.4	38.2	20.6	60.7	8.3	25.2	3.2	42.5	95.5	1.0	95.8	24.6	27.5	34.0	7.1	7.6	29.0	96.2	36.7	32.5	38.2
Ours	82.7	46.4	6.4	39.7	51.1	69.4	15.3	29.6	3.0	40.0	89.4	1.2	96.8	35.6	29.2	28.4	6.8	10.6	32.3	96.9	**40.5**	**35.6**	**40.4**
Ours + PointPainting	79.8	30.7	16.2	42.5	47.6	53.4	8.1	21.7	3.9	32.8	82.3	0.0	95.6	34.2	19.6	25.3	11.9	19.7	30.0	96.4	37.6	28.9	34.8
w/o HHIO	79.5	15.2	17.4	32.9	31.6	56.6	7.1	26.1	1.8	35.0	92.8	0.6	95.8	22.0	26.7	30.2	9.5	19.0	29.0	97.1	36.3	25.0	31.6
Ours + LocalFusion	81.8	46.8	2.0	46.2	36.8	74.7	13.2	28.5	0.4	37.3	93.8	2.3	96.5	35.2	37.0	27.8	8.6	9.9	26.0	96.5	40.1	36.9	42.0
w/o HHIO	80.7	39.3	2.3	41.9	26.6	73.1	13.9	23.2	2.2	35.4	92.2	0.0	94.4	24.5	29.7	31.4	7.3	13.6	36.1	95.9	38.2	36.6	41.6

Table 3. Semantic segmentation results on BEHAVE dataset.

	person	backpack	basketball	boxlarge	boxlong	boxmedium	boxsmall	boxtiny	chairblack	chairwood	keyboard	monitor	container	stool	suitcase	tablesmall	tablesquare	toolbox	trashbin	yogaball	yogamat	mIOU
SoftGroup[55]	96.8	71.8	54.9	83.7	54.4	56.3	40.4	21.9	85.6	82.4	27.1	67.3	71.7	83.5	77.7	82.5	90.5	44.5	64.7	89.0	70.5	67.5
Ours	97.0	72.4	61.3	86.6	62.2	57.3	45.6	33.4	87.7	83.3	30.8	72.1	73.8	84.2	76.0	86.6	91.9	49.2	66.9	89.0	76.4	**70.7**

some objects are relatively small and closely located to the person, causing overlapping or stitching points with humans and bringing difficulties in distinguishing from the person. To tackle these problems, we propose a Human-Human-Object Interaction(HHOI) module, shown in Figure 3. The model first extracts the human-human interaction feature with attention strategy so that humans can be more accurately recognized even with partial point cloud in occluded scenes. Then, it uses human-centric features to guide the network automatically to learn a weighted feature to pay attention to interactive objects, which can benefit capturing fine-grained semantic information.

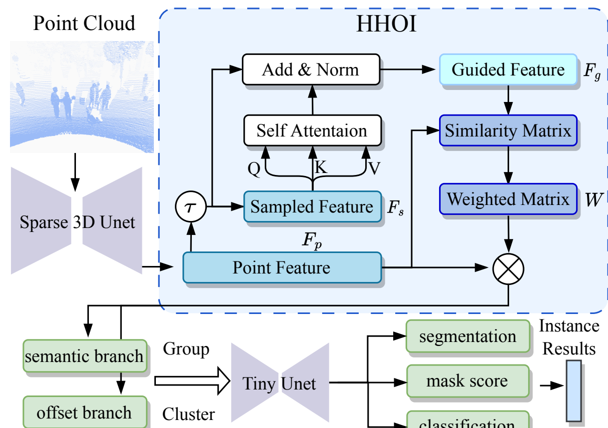

Figure 3. The architecture of our segmentation method. Especially, the HHOI module extracts the correlation within different persons and the human-object relationships, which can benefit the point-wise and instance-wise classification.

5.1.1 Human-Human-Object-Interaction Module

As shown in Figure 3, we utilize a sparse 3D Unet to get D dimensional point feature $F_p \in \mathcal{R}^{N\times D}$. Then, human-human interacted features are extracted through a transformer mechanism. We get the semantic score $Y = softmax(MLP(F_p)) = \{y_{i,c}\}^{N\times C}$ for each point, where C is the class number. And then we select M points with the confidence of belonging to person class higher than the threshold τ. We further apply the triplet Q, K, V attention layer to extract correlations among different sampled person features F_s and obtain the final human-guided feature:

$$f_{attention} = softmax(\frac{QK^T}{\sqrt{D}})V,$$

$$F_g = LN(f_{attention} + FFN(f_{attention})),$$

where LN is layer normalization and FFN is the feed-forward neural network [54]. Then, we use human-guided feature to extract human-object interaction for fine-grained object segmentation. The similarity weighted matrix $W = softmax(F_p F_g^T)$ is computed to enhance the features of objects that people interact with. We multiply the weighted matrix with point features to obtain the final weighted features. In this way, the model adaptively learns human-related representations and enhances the object feature with the guidance of high-confidence human features.

5.1.2 Point-wise Prediction and Refinement

Taking the weighted features as input, the semantic branch and offset branch apply two-layer MLP and output the semantic scores $S \in \mathcal{R}^{N\times K}$ and offset vectors $O \in \mathcal{R}^{N\times K}$ from the point to the instance center, respectively. The weighted cross-entropy loss $\mathcal{L}_{\text{semantic}}$ and L_1 regression loss $\mathcal{L}_{\text{offset}}$ are used to train the semantic and offset branches. After that, we follow the refinement stage in SoftGroup [55], where point-level proposals are fed into a tiny-unet to predict classification scores, instance masks, and mask scores to generate the final instance results. Specifically, the classification branch predicts the category scores c_k for each instance. The segmentation branch utilizes a point-wise MLP to predict an instance mask m_k for each instance proposal. Mask scoring branch estimates the IoU between the pre-

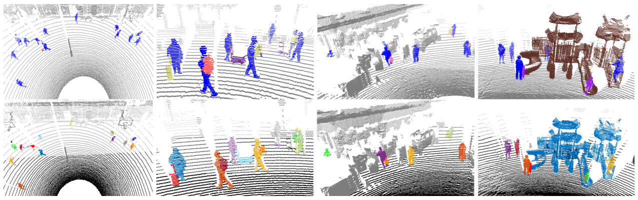

Figure 4. The visualization of semantic (first row) and instance (second row) segmentation results of our method on HuCenLife.

dicted mask and the ground truth for each instance. We train each branch with cross-entropy loss $\mathcal{L}_{\text{class}}$, binary cross-entropy loss $\mathcal{L}_{\text{mask}}$, and l_2 regression loss $\mathcal{L}_{\text{mask score}}$. And the total loss is the sum of all above losses.

5.2. Experiments

5.2.1 Baselines and Evaluation Metrics

Previous 3D instance segmentation works can be divided into LiDAR-based methods and RGB-D-based methods. For the former, we compare with current SOTA method DSNet [26] of both voxel-division version and cylinder-division version. For the latter, we select current SOTA approaches DKnet [58] and SoftGroup [55] for comparison.

We utilize mean IoU (mIoU) to evaluate the quality of the semantic segmentation. For instance segmentation, we report AP50 and AP25 which denote the scores with IoU thresholds of 50% and 25%, respectively.

5.2.2 Results

Comparison on HuCenLife dataset. We compare the results of our proposed method with baseline methods in Table 2. DSNet does not get satisfactory results, mainly because it focuses on traffic scenarios, while the span of object scale is much larger in human-centric scenarios. SoftGroup is better than outdoor methods because it has a refinement stage for recognizing small objects. Our method performs best due to the use of interaction information.

Comparison on BEHAVE dataset. To further evaluate the generalization capability of human-object interaction scenes, we also conduct experiments for semantic segmentation on BEHAVE [5] dataset in Table 3. BEHAVE dataset is a human-object interaction dataset, which is collected in indoor scenarios and provides RGB-D frames and 3D SMPL. To adapt it to our task, we generate the point cloud and segmentation label from RGB-D images and segmented masks. There is only single person with single object per frame and the total number of the object categories is 20. We follow the official protocol of dataset splitting. Our method still outperforms the best baseline method SoftGroup by 2.8% in mIOU.

Sensor-fusion-based 3D segmentation. Because our dataset also contains image data, we also provide LiDAR-Camera sensor-fusion baselines based on our method in Table 2 to facilitate further research. PointPainting appends the raw LiDAR point with corresponding RGB color according to calibration matrix. LocalFusion concatenates high-dimensional image feature to the corresponding high dimensional point semantic feature. And our HHOI module has consistently improved the performance on various fusion strategies, validating its generalization ability.

Table 4. Person-only 3D detection results on HuCenLife.

Methods	AP(0.25)	AP(0.5)	AP(1.0)	mAP
CenterPoint[65]	61.8	68.7	70.3	66.9
STCrowd[11]	61.8	71.6	73.4	68.9
TED[57]	51.0	53.3	54.1	52.8
CenterFormer[75]	73.0	80.1	81.4	78.2

Table 5. Full-category 3D detection results (AP) on HuCenLife. We only select six types of objects for demonstration.

Methods	motorbike	box	cart	scooter	backpack	object in hand
CenterPoint[65]	13.4	17.1	20.9	43.4	4.2	8.4
STCrowd[11]	5.4	14.4	25.3	48.7	4.5	13.5
CenterFormer[75]	3.8	16.2	24.2	44.4	2.6	12.5

6. Human-centric 3D Detection

LiDAR point cloud-based 3D detection is well-studied in recent years, driven by autonomous driving. It provides critical information of obstacles for the motion planning of robots to guarantee the safety. Specifically, the input for 3D detection is the point cloud P and the output is predicted bounding boxes with 7 dimensions (x,y,z,w,l,h,r), consisting of the 3D position in LiDAR coordinate system, the size of bounding box, and the rotation. In this section, we provide benchmarks for the 3D detection task on HuCenLife by evaluating current state-of-the-art methods and give discussion on the research of human-centric 3D detection.

6.1. Baselines and Evaluation Metrics

We choose four representative works and test their performance on our dataset. CenterPoint [65] is a popular anchor-free detector and based on it, STCrowd [11] aims at solving dense crowd scenarios. By means of the transformer mechanism, TED [57] and CenterFormer [75] achieve impressive performance recently. Following [11, 6], we use Average Precision (AP) with 3D center distance thresholds D = $\{0.25, 0.5, 1\}$ meters as the evaluation metric. Then mean Average Precision (mAP) is obtained by averaging AP.

6.2. Results and Discussion

We conduct experiments on two settings, including **person-only 3D detection** in Table 4 and **full-category 3D detection** in Table 5. These baseline methods are designed for large-scale traffic scenarios, which perform limited on human-centric scenarios, especially for detecting small objects. We conclude with three main challenges for conduct-

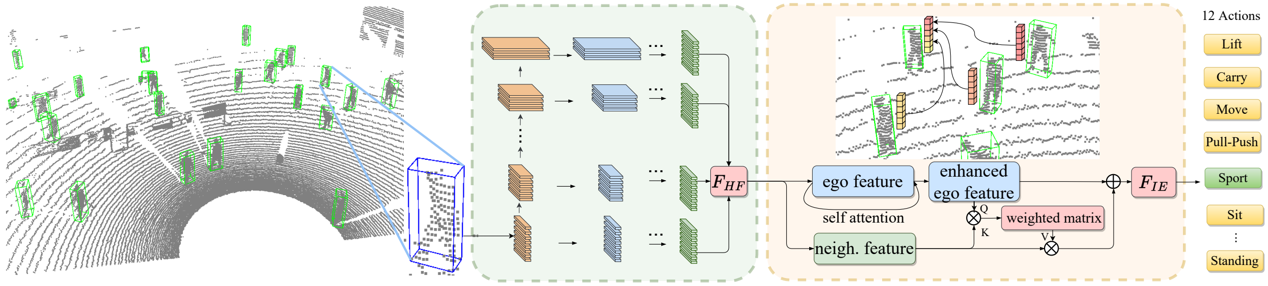

Figure 5. Pipeline of our method for human-centric action recognition. We first utilize 3D detector to obtain a set of bounding boxes of persons. Then, for each person, we extract multi-resolution features and get a hierarchical fusion feature F_{HF}. Next, we leverage the relationship with neighbors to enhance the ego-feature and obtain a comprehensive feature F_{IE} for the final action classification.

ing 3D detection in human-centric scenarios. First, people usually have different poses in different actions, such as crouching, sitting, waving, etc., and such diverse body poses cause distinct sizes of bounding box. Second, there are many relatively small objects in scenes, bringing difficulties to balance the accuracy of fine-grained detection and the efficiency of large-scale scene data processing. Third, multi-objects may locate at different heights in the same place, such as in complex scenarios of escalator and slide, leading to larger dimension of feature recognition. Previous methods using BEV feature map will miss details and transformer-based methods have horrible cost. Therefore, there is a lot of room for the 3D detection research in human-centric scenes, while our dataset can offer a good platform for it.

7. Human-centric Action Recognition

Previous works for action recognition are based on 2D images or videos and they only need to give one label for one scene. We introduce the 3D action recognition task in large-scale human-centric scenarios, which aims to detect all persons in the scene and provide corresponding action types. 3D action recognition task is significant for fine-grained scene understanding and can benefit the development of intelligent surveillance and collaborative robots. To our knowledge, we are the first to propose the related dataset and solutions for the new task.

7.1. Method

Our 3D action recognition method is in a two-stage manner based on the input of LiDAR point cloud, as shown in Figure 5. Considering that some human actions are related to adjacent interactive objects, after obtaining individual bounding box by 3D detector, we enlarge the box to crop more points related to the person for the following fine-grained feature extraction. Especially, we leverage a Hierarchical Point Feature Extraction module to pay attention to multi-scale objects and get multi-level features. Moreover, we design an Ego-Neighbour Feature Interaction (ENFI) module to make use of the relationship among the ego-person and neighbors to help forecast social actions.

7.1.1 Hierarchical Point Feature Extraction

To capture both global features and local features with dynamically changing receptive fields, we use R parallel branches to extract multi-resolution features. Serial Set Abstractions [39] are applied to process the features of different scales, where each branch undergoes L times with fixed sampling cores and branch-specific sampling range. Finally, these features are up-sampled to the same dimension and fused together with pooling to generate the hierarchical fusion feature F_{HF}.

7.1.2 Ego-Neighbour Feature Interaction

Like Figure 5 shows, we first enhance the ego person feature by self-attention and get F_{ego}. Then, we select features of k neighbours around the target as K_{neigh} and V_{neigh} and take the ego-feature as queries Q_{ego}. The distances from neighbours to the target are used for position encoding. We apply cross-attention to extract the ego-neighbour interaction information and gain the final interaction enhanced ego feature by $F_{IE} = F_{ego} \bigoplus \text{CrossAttention}(Q_{ego}, K_{neigh}, V_{neigh})$, where $\bigoplus$ denotes concatenation. In this way, we model the relationships of a group to benefit the social action recognition.

Table 6. Comparison results of action recognition on HuCenLife. All methods are based on the same 3D detector for fair evaluation.

Methods	mAP	mRecall	mPrecision
Baseline	7.3	14.6	19.9
+ ViT[17]	9.4	23.1	19.9
+ PVT[68]	13.2	30.5	19.8
+ PointNet[39]	8.4	26.3	15.5
+ PointNet++[40]	15.6	34.2	22.7
+ PointMLP[34]	11.3	28.0	19.4
+ PointNeXt[41]	15.0	33.0	21.2
Ours	**21.0**	**40.0**	**26.9**
Ours(w/o ENFI)	15.4	37.1	24.7

7.2. Experiments

7.2.1 Baselines and Evaluation Metrics

We take pre-trained CenterPoint as the 3D Detector for all the experiments for fair comparison in this section. Because no existing methods can be directly used for solving the new 3D action recognition task. As Table 6 shows, we provide benchmarks and comparisons from four aspects. The first is to directly adapt the 3D detector to predict multi-class persons with different action labels, which is the "Baseline" in Table 6. The second is to add a feature extractor for cropped individual point cloud for the second-stage action classification, and we tried several popular point-feature extractors, including PVT, PointNet, PointNet++, PointMLP, and PointNext. In particular, to verify the performance of input modalities, we also use ViT to extract image features for image-based action recognition by projecting the 3D bounding box to calibrated images. At last, we provide the results of our solution with ablation for ENFI module.

We use the mean Average Precision (mAP) obtained by averaging AP through thresholds $D = \{0.25, 0.5, 1\}$ and classes to evaluate the performance.

$$\mathrm{mAP} = \frac{1}{|\mathbb{C}||\mathbb{D}|} \sum_{c \in \mathbb{C}} \sum_{d \in \mathbb{D}} \mathrm{AP}_{c,d}$$

where $|\mathbb{C}|$ is the number of action category. In addition, we also utilize Mean Recall (mRecall) and Mean Precision (mPrecision) by averaging recall and precision through thresholds and classes.

7.2.2 Results and Discussion

We show the overall performance in Table 6, and detailed evaluation values of all categories of actions and visualization results are in the supplementary material. It can be seen from the results that our method outperforms others with an obvious margin, mainly due to the multi-level feature extraction and multi-person interaction modeling, which are more suitable for understanding human-centric complex scenarios. However, our method has its own limitations and there are several potential improvement directions. First, current two-stage framework strongly relies on the detector performance and the one-stage method for action recognition in large-scale scenes is worth exploring. Moreover, human action is time-dependent and how to extract valuable temporal information in consecutive data to eliminate the ambiguity of actions is also promising.

8. More Tasks on HuCenLife

In this paper, we provide benchmarks on HuCenLife for three main tasks, including 3D segmentation, 3D detection, and action recognition in human-centric scenarios. However, benefiting from the rich annotations in HuCenLife dataset, there are many other tasks deserving explored.

8.1. Human-Object Interaction Detection

Recently, the task of Human-Object Interaction (HOI) detection [70, 66] attracts more and more attention, which targets for detecting the person and the interacted object and meanwhile classifying the interaction category. Current studies and datasets are limited to the interaction between single person and single object in one scene and they are all based on the image modality. 3D HOI tasks in large-scale free environments with multiple persons and multiple objects can be formulated and evaluated on HuCenLife.

8.2. Tracking and Trajectory Prediction

HuCenLife contains sequential frames of data with the tracking ID annotation for all instances, which can facilitate the time-related tasks, such as 3D tracking [73, 72] and trajectory prediction [35, 16]. It is challenging for these tasks due to the occlusions in crowded scenes, but it is significant to study consecutive behaviors and interactions in real world to provide valuable guidance for robots.

8.3. 3D Scene Generation

With the success of Diffusion model [25] in image generation, many works try to achieve high-quality 3D data generation for single objects [33] or scenes [77]. HuCenLife provides rich material for daily-life scenarios, and it is interesting to generate more human-centric scene data with semantic information to facilitate learning-based methods.

8.4. Multi-modal Feature Fusion

Apart from point cloud, HuCenLife also provides corresponding images. The complementary information of multi-modal features will definitely benefit all tasks mentioned above, which deserves in-depth research.

9. Conclusion

We fully discuss the challenges, significance, and potential research directions of 3D human-centric scene understanding in this paper. Specifically, we propose the first related large-scale dataset with rich fine-grained annotations, which can facilitate the research for many 3D tasks and has the potential to boost the development of assistive robots, surveillance, etc. Moreover, we provide benchmarks for various tasks and propose novel methods for human-centric 3D segmentation and human-centric action recognition to facilitate further research.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, pages 1534–1543, 2016. 2, 4

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 2, 3

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019. 2, 4

[4] Alex Bewley, Pei Sun, Thomas Mensink, Dragomir Anguelov, and Cristian Sminchisescu. Range conditioned dilated convolutions for scale invariant 3d object detection. In *Conference on Robot Learning*, pages 627–641. PMLR, 2021. 3

[5] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *CVPR*, pages 15935–15946, 2022. 2, 4, 6

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 2, 4, 6

[7] Chen Chen, Zhe Chen, Jing Zhang, and Dacheng Tao. Sasa: Semantics-augmented set abstraction for point-based 3d object detection. In *AAAI*, number 1, pages 221–229, 2022. 3

[8] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *ICCV*, pages 15467–15476, 2021. 3

[9] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In *ICCV*, pages 9775–9784, 2019. 3

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 2

[11] Peishan Cong, Xinge Zhu, Feng Qiao, Yiming Ren, Xidong Peng, Yuenan Hou, Lan Xu, Ruigang Yang, Dinesh Manocha, and Yuexin Ma. Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes. In *CVPR*, pages 19608–19617, 2022. 1, 2, 4, 6

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 2, 4

[13] Yudi Dai, Yi Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. *CVPR*, pages 6782–6792, 2022. 1

[14] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, volume 35, pages 1201–1209, 2021. 3

[15] Shengheng Deng, Zhihao Liang, Lin Sun, and Kui Jia. Vista: Boosting 3d object detection via dual cross-view spatial attention. In *CVPR*, pages 8438–8447. IEEE, 2022. 3

[16] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pages 203–212. PMLR, 2022. 8

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 7

[18] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *CVPR*, pages 9031–9040, 2020. 3

[19] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Rangedet: In defense of range view for lidar-based 3d object detection. In *ICCV*, pages 2898–2907. IEEE, 2021. 3

[20] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 3

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 CVPR*, pages 3354–3361. IEEE, 2012. 2, 4

[22] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 2

[23] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *CVPR*, pages 2940–2949, 2020. 3

[24] Tong He, Chunhua Shen, and Anton Van Den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *CVPR*, pages 354–363, 2021. 3

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 8

[26] Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *CVPR*, pages 13090–13099, 2021. 2, 5, 6

[27] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, pages 4867–4876, 2020. 3

[28] Tianyuan Jiang, Nan Song, Huanyu Liu, Ruihao Yin, Ye Gong, and Jian Yao. Vic-net: voxelization information compensation network for point cloud 3d object detection. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13408–13414. IEEE, 2021. 3

[29] Junho Koh, Junhyung Lee, Youngwoo Lee, Jaekyum Kim, and Jun Won Choi. Mgtanet: Encoding sequential lidar

points using long short-term motion-guided temporal attention for 3d object detection, 2022. 3
[30] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 3
[31] Jialian Li and etc. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *CVPR*, pages 20502–20512, 2022. 1
[32] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 3
[33] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pages 2837–2845, 2021. 8
[34] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 7
[35] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, volume 33, pages 6120–6127, 2019. 8
[36] WU Peng, GU Lipeng, YAN Xuefeng, XIE Haoran, Fu Lee WANG, Gary CHENG, and WEI Mingqiang. Pv-rcnn++: semantical point-voxel feature interaction for 3d object detection. *Visual Computer*, 2022. 3
[37] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning. *arXiv preprint arXiv:2212.03229*, 2022. 2, 3
[38] Charles R Qi, Or Litany, Kaiming He, and Leonidas Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9276–9285. IEEE. 3
[39] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 2, 7
[40] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 7
[41] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *arXiv preprint arXiv:2206.04670*, 2022. 7
[42] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1
[43] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 3
[44] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 2, 3, 4
[45] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, and etc Shi, Jianping. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10526–10535. IEEE Computer Society, 2020. 3
[46] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779. IEEE Computer Society, 2019. 3
[47] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015. 3, 4
[48] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 3
[49] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 2
[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020. 2, 4
[51] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *CVPR*, pages 5725–5734, 2021. 3
[52] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, pages 6411–6420, 2019. 2
[53] Zhi Tian, Xiangxiang Chu, Xiaoming Wang, Xiaolin Wei, and Chunhua Shen. Fully convolutional one-stage 3d object detection on lidar range images. In *Advances in Neural Information Processing Systems*. 3
[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
[55] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *CVPR*, pages 2708–2717, 2022. 3, 5, 6
[56] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *ICRA*, pages 1887–1893. IEEE, 2018. 3
[57] Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-equivariant 3d object detection for autonomous driving. *arXiv preprint arXiv:2211.11962*, 2022. 3, 6
[58] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *ECCV*, pages 235–252. Springer, 2022. 3, 5, 6

[59] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, pages 3333–3343, 2022. 2, 3

[60] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3

[61] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *Advances in neural information processing systems*, 32, 2019. 3

[62] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *CVPR*, pages 11037–11045. IEEE, 2020. 3

[63] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*, pages 1951–1960, 2019. 3

[64] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *CVPR*, pages 3947–3956, 2019. 3

[65] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 3, 6

[66] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *AAAI*, volume 36, pages 3206–3214, 2022. 8

[67] Wang Yue, Sun Yongbin, Liu Ziwei, Sanjay E Sarma, and Michael M Bronstein. Dynamic graph cnn for learning on point clouds. *TOG*, 38(5), 2019. 2

[68] Cheng Zhang, Haocheng Wan, Xinyi Shen, and Zizhao Wu. Pvt: Point-voxel transformer for point cloud learning. *International Journal of Intelligent Systems*, 37(12):11985–12008, 2022. 7

[69] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *CVPR*, pages 18931–18940. IEEE, 2022. 3

[70] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, pages 19548–19557, 2022. 8

[71] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *CVPR*, pages 9601–9610, 2020. 2

[72] Chaoda Zheng, Xu Yan, Haiming Zhang, Baoyuan Wang, Shenghui Cheng, Shuguang Cui, and Zhen Li. Beyond 3d siamese tracking: A motion-centric paradigm for 3d single object tracking in point clouds. In *CVPR*, pages 8111–8120, 2022. 8

[73] Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Haiyu Zhao, and Shijian Lu. Pttr: Relational 3d point cloud object tracking with transformer. In *2022 CVPR*, pages 8521–8530. IEEE Computer Society, 2022. 8

[74] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pages 4490–4499. IEEE Computer Society, 2018. 3

[75] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *ECCV*, pages 496–513. Springer, 2022. 3, 6

[76] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, pages 9939–9948, 2021. 2

[77] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point clouds. In *ECCV*, pages 17–35. Springer, 2022. 8