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

Yiteng Xu\textsuperscript{1,}\textsuperscript{*}, Peishan Cong\textsuperscript{1,}\textsuperscript{*}, Yichen Yao\textsuperscript{1,}\textsuperscript{*}, Runnan Chen\textsuperscript{2}, Yuenan Hou\textsuperscript{3}, Xinge Zhu\textsuperscript{4}, Xuming He\textsuperscript{1}, Jingyi Yu\textsuperscript{1}, Yuexin Ma\textsuperscript{1,}\textsuperscript{†}

\textsuperscript{1} ShanghaiTech University \textsuperscript{2} The University of Hong Kong
\textsuperscript{3} Shanghai AI Laboratory \textsuperscript{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 \texttt{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-
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. Have [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-based methods [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.

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° × 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
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-centric 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 \( P \in \mathbb{R}^{N \times 4} \), 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,
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.

### 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 \( F_p \in \mathbb{R}^{N \times D} \). Then, human-interacted features are extracted through a transformer mechanism. We get the semantic score \( Y = \text{softmax}(\text{MLP}(F_p)) = \{y_{c,i}\}_{N \times C} \) 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 \( \tau \). 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:

\[
    f_{\text{attention}} = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V,
\]

\[
    F_g = \text{LN}(f_{\text{attention}} + \text{FFN}(f_{\text{attention}})),
\]

where \( \text{LN} \) is layer normalization and \( \text{FFN} \) 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 \( W = \text{softmax}(F_gF_g^T) \) 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 \( S \in \mathbb{R}^{N \times K} \) and offset vectors \( O \in \mathbb{R}^{N \times K} \) from the point to the instance center, respectively. The weight 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-
dicted mask and the ground truth for each instance. We train each branch with cross-entropy loss $L_{\text{class}}$, binary cross-entropy loss $L_{\text{mask}}$, and $l_2$ regression loss $L_{\text{mask score}}$. 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>
<thead>
<tr>
<th>Methods</th>
<th>motorbike</th>
<th>box</th>
<th>cart</th>
<th>scooter</th>
<th>backpack</th>
<th>object in hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>CenterPoint[65]</td>
<td>13.4</td>
<td>17.1</td>
<td>20.9</td>
<td>43.4</td>
<td>4.2</td>
<td>8.4</td>
</tr>
<tr>
<td>CenterFormer[75]</td>
<td>3.8</td>
<td>16.2</td>
<td>24.2</td>
<td>44.4</td>
<td>2.6</td>
<td>12.3</td>
</tr>
</tbody>
</table>

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.

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 $F_{IE} = F_{ego} \oplus \text{CrossAttention}(Q_{ego}, K_{neigh}, V_{neigh})$, where $\oplus$ 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.

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

$$mAP = \frac{1}{|C||D|} \sum_{c \in C} \sum_{d \in D} AP_{c,d}$$

where $|C|$ 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


[59] Shaochun Han, Jiacheng Wang, Zhi Tian, Xiachen Wang, Chaoxu Guo, Shaoshuai Shi, and etc. Mask3d for 3D semantic point-voxel feature interaction for 3D object detection, 2022.


[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


