HUNTER: Unsupervised Human-centric 3D Detection via Transferring Knowledge from Synthetic Instances to Real Scenes

Yichen Yao[1] , Zimo Jiang[1,2,3], Yujing Sun[4], Zhencai Zhu[3], Xinge Zhu[5], Runnan Chen[4], Yuexin Ma[1,*]
[1] ShanghaiTech University [2] University of Chinese Academy of Sciences
[3] Innovation Academy for Microsatellites, Chinese Academy of Sciences
[4] The University of Hong Kong [5] The Chinese University of Hong Kong
{yaoych2023,mayuexin}@shanghaitech.edu.cn

Abstract

Human-centric 3D scene understanding has recently drawn increasing attention, driven by its critical impact on robotics. However, human-centric real-life scenarios are extremely diverse and complicated, and humans have intricate motions and interactions. With limited labeled data, supervised methods are difficult to generalize to general scenarios, hindering real-life applications. Mimicking human intelligence, we propose an unsupervised 3D detection method for human-centric scenarios by transferring the knowledge from synthetic human instances to real scenes. To bridge the gap between the distinct data representations and feature distributions of synthetic models and real point clouds, we introduce novel modules for effective instance-to-scene representation transfer and synthetic-to-real feature alignment. Remarkably, our method exhibits superior performance compared to current state-of-the-art techniques, achieving 87.8% improvement in mAP and closely approaching the performance of fully supervised methods (62.15 mAP vs. 69.02 mAP) on HuCenLife Dataset.

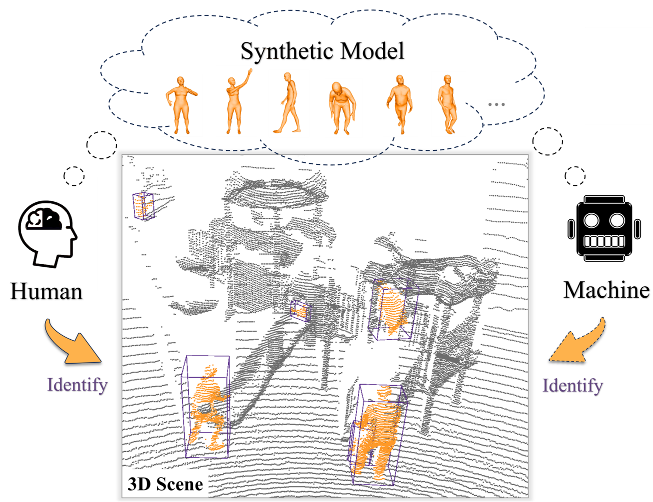


Figure 1. Human has the ability to identify objects in 3D scenes, relying merely on their understanding of the objects' shapes and sizes. We aspire for machines to possess the capability to perform 3D perception solely based on synthetic models, independent of any scene-level annotations.

1. Introduction

The field of 3D scene understanding in human-centric scenarios has garnered increasing attention in recent years, owing to its pivotal role in the advancement of research on humanoid robots, assistive robots, and human-robot collaboration. To navigate safely within 3D space and effectively interact with humans, it is crucial for robots to possess the capability to accurately perceive and localize individuals. Consequently, some LiDAR-based human-centric 3D perception datasets and methods [10, 41] have been proposed in recent years, aimed at propelling progress in this domain.

In contrast to the domain of traffic perception for autonomous driving [1, 3, 49, 50], the realm of human-centric perception presents a significantly more formidable challenge. Unlike the relative regular object distribution and background context in traffic scenarios, real-life human-centric settings encompass a vast spectrum of indoor and outdoor environments with intricate and diverse backgrounds. Moreover, unlike rigid vehicles, humans exhibit ever-changing poses, movements, and trajectories, and engage in multifaceted interactions with objects and their surroundings. All these factors cause difficulties for data acquisition and data annotation, concurrently posing challenges to the precision and generalization capacity of perception methodologies. Therefore, the research for effective unsupervised methods of human-centric 3D perception

*Corresponding author. This work was supported by NSFC (No.62206173), Natural Science Foundation of Shanghai (No.22dz1201900), Shanghai Sailing Program (No.22YF1428700), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

becomes necessary and imperative.

Current mainstream methods for unsupervised 3D detection can be categorized into two paradigms. One [23, 36, 45] relies on motion information, such as scene flow, to discriminate between foreground and background elements. However, these methods encounter limitations when confronted with static objects. The other [46, 47] harnesses point clustering algorithms to derive pseudo-labels for subsequent iterative self-training. Nevertheless, these pseudo-labels tend to exhibit poor quality in human-centric scenarios and making the self-learning process worse and worse. That is because humans usually have sparse points captured by LiDAR, and humans may stay very close with others, all presenting considerable challenges for clustering algorithms to distinguish human instances effectively. Furthermore, for all these methods, they cannot furnish semantic information for detected objects, necessitating additional classifiers to identify the human class.

Indeed, humans exhibit an impressive ability to detect objects in 3D spaces based solely on their knowledge of object shapes and sizes, as Fig. 1 shows. Since we can generate arbitrary human instances by the parametric model [20], this prompts us to ponder the following question: "*Can our AI method detect human instances in 3D scenes without any scene-level annotations, merely relying on synthetic human models?*" To achieve this, we must address three pivotal challenges: the unification of disparate data representations of mesh models and LiDAR point clouds; the alignment of dissimilar feature distributions of synthetic humans and real humans; and the exploitation of prior knowledge specific to the human body to enhance perceptual acuity.

In this paper, we propose a novel method, named HUNTER, for unsupervised **HU**man-centric 3D detectio**N** via **T**ransferring knowledge from synth**E**tic instances to **R**eal scenes. To tackle three crucial issues mentioned above, we design three corresponding stages in our method, including **instance-to-scene representation transfer**, **synthetic-to-real feature alignment**, and **fine-grained perception enhancement**. Firstly, we insert synthetic human models into 3D scenes and employ range-view projection to transform the mesh representation into LiDAR point clouds, which can imitate the point distribution patterns in the insertion place and preserve correct partial point clouds caused by occlusions. Utilizing labels associated with synthetic humans, we train our 3D detector to perceive pseudo humans in the scene. Secondly, to enhance the generalize ability of our detector to real humans, we perform feature alignment between synthetic and genuine instances. Notably, we employ a filtering strategy to select high-quality real samples for effective feature alignment. Third, recognizing the specific structural constraints of the human body, we employ the body skeleton as an additional supervisory signal. This approach enhances fine-grained feature learning for humans and alleviates the challenge of detecting incomplete human point clouds resulting from occlusions. Finally, we integrate these three components into a comprehensive self-training framework to achieve unsupervised learning for human-centric 3D detection.

To evaluate the effectiveness of our method, we conduct experiments on two large-scale 3D datasets focusing on human-centric scenarios, including STCrowd [10] and HuCenLife [41]. Results show that HUNTER outperforms current state-of-the-art methods with a large margin (87.8% improvement) and performs close to fully supervised performance (62.15 mAP v.s. 69.02 mAP) on HuCenLife. Our contribution can be summarized as follows:

- We propose the first unsupervised 3D detection method for human-centric scenarios, which is significant for the development of robotics in real-life applications.
- We present a novel solution by transferring the knowledge from synthetic human models to real 3D scenes.
- Our method demonstrates exceptional performance, achieving SOTA results on open datasets and closely rivaling fully supervised approaches.

2. Related Work

2.1. LiDAR-based 3D Detection

LiDAR-based 3D detection [16, 38, 42, 44, 48–50] is fundamental for robots to understand large-scale scenes, so that robots can navigate and conduct tasks in 3D space safely and effectively. This task has been well studied for many years in traffic scenes [1, 3, 22, 32] and boosted the development of autonomous driving. In recent years, human-centric scene understanding [12, 17, 30] in 3D large-scale scenarios is attracting increasing attention, which is significant for human-robot interaction and human-robot collaboration. Several human-centric datasets have been proposed, including STCrowd [10], focusing on the pedestrian detection task in crowded scenarios, and HuCenLife [41], emphasizing perceiving humans with varied poses and activities in diverse daily-life scenarios. However, current detection methods suitable for human-centric scenarios are all supervised, which requires amounts of annotations and has poor generalization ability for novel scenes.

2.2. Unsupervised 3D Object Discovery

The task of unsupervised object discovery [33, 33, 35] aims to recognize or localize objects without relying on costly annotated training data. In the realm of 3D point clouds [2, 15, 28, 31], researchers mainly using the geometric or dynamic properties to distinguish objects and backgrounds. Both [23] and [36] utilize the motion information, scene flow, to detect moving objects. [45] requires multiple traversals over the same location to filter dynamic objects. However, theses methods could not perceive static objects. Another mainstream approaches [40, 46, 47] leverage cluster algorithms [13] to generate initial pseudo-labels for instances and iteratively self-train the model to improve the label quality. However, pseudo-labels often exhibit poor quality in human-centric scenarios, making the self-learning process increasingly worse. This is because humans, particularly those who are far away from LiDAR, typically have only a few sparse points, and in daily life scenarios, humans may be in close proximity to other objects or instances, making it challenging for clustering algorithms to distinguish human instances effectively. Furthermore, a common limitation of existing methods is their inability to provide semantic information for detected objects, necessitating the need for an additional classifier to identify the human class. In contrast, our pseudo-labels for synthetic humans naturally have high quality and semantics.

2.3. Transfer Learning in 3D

In order to improve the generalization capability of network under limited data, transfer learning has been widely employed in 3D perception tasks, which contains several categories of paradigms, such as pre-training [18, 39, 43], domain adaptation [26, 27], weak supervision [11], zero-shot/open-vocabulary [5–7, 21, 25], etc. There are some recent works [4, 8, 29] also aim to transfer the knowledge of synthetic models to real scenes, which mix 3D objects with randomized layouts to synthesize scenes. These methods are suitable for indoor scenarios with regular layouts and simple background. [24] proposes unsupervised 3D perception by distilling knowledge from 2D vision-language pre-training. It applies to outdoor autonomous driving scenarios. However, they are not applicable for LiDAR-based large-scale human-centric scenes, where LiDAR point patterns differs across distances, the layouts of real-life scenes are dramatically diverse, and backgrounds are always changing and much more complex.

3. Methods

Human-centric 3D Detection aims to identify humans in 3D scenes. In this paper, we propose HUNTER, for unsupervised 3D human detection via transferring knowledge from synthetic instances to real scenes. Our method mainly contains three stages, including instance-to-scene representation transfer, synthetic-to-real feature alignment, and fine-grained perception enhancement, as shown in Fig. 2. For the first stage, we exploit the inherent class-aware attributes of simulated human instances to generate pseudo-labels for real humans. In the second stage, we utilize bi-directional multi-object tracking to filter out low-quality pseudo-labels. Then, we conduct feature alignment to bridge the gap between synthetic humans and real-world captured humans. For the third stage, we integrate the human body skeleton as supervision to enhance fine-grained feature learning, which can tackle more challenging cases such as human with severe occlusions. We present the details in the following.

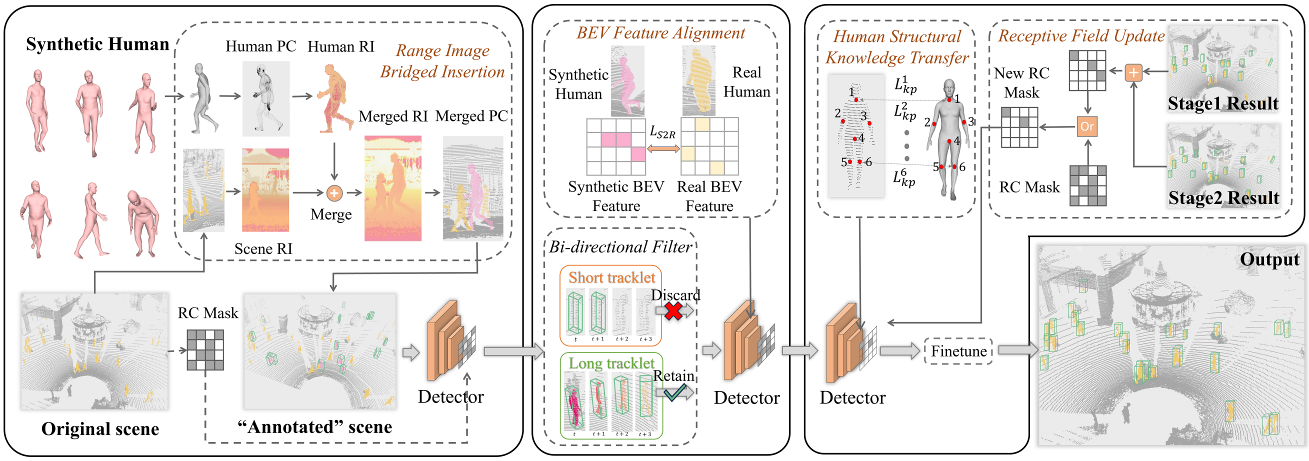
3.1. Instance-to-Scene Representation Transfer

Directly inserting synthetic human instances into scenes is an intuitive solution to create annotated data. However, where to insert and how to insert are two critical issues. We first utilize ground-guided synthetic human insertion to make the insert positions approach the real distribution of humans in scenes. Then, considering that synthetic human instances are represented in mesh form, which dramatically differs from LiDAR point cloud, we propose range image-bridged point cloud generation to achieve the representation transfer. Moreover, we employ mask-constrained receptive field control to guide the model's training. Finally, class-aware pseudo-labels for real humans are generated by model inferring.

Ground-guided synthetic human insertion. Following the common sense of the real human locations in the scene, we randomly place the synthetic human instance on the ground. Specifically, we adopt the approach in MODEST [45] and employ RANSAC [14] to obtain the ground point cloud data. Afterwards, we place selected synthetic human instances on randomly selected ground locations. As shown in Fig. 2, we adjust the translation matrix of human instance to let its lowest point coincide with ground point. To ensure the inserted human instances obey the real LiDAR point distribution, we simulate a LiDAR system [9] to obtain the sparse human point cloud.

Range image bridged point cloud generation. However, the simulated LiDAR point cloud only accounts for self-occlusion. To generate scenes that adhere to the view-dependent property of LiDAR point cloud, we simulate the external occlusions by other instances or objects in the scene for inserted humans. Specifically, we utilize the range image $I^{H*W*(N+1)}$ to achieve this, where H and W represents the vertical and horizontal resolution of LiDAR, and N represent the dimension of point cloud (e.g. x, y, z). Due to LiDAR's limited measuring range, the image has empty grids. Thus, one more dimension is added to represent whether points are present in the grid. If multiple points in the real scene fall into the same grid, LiDAR only records the closest point due to the physical nature of light, which propagates in a straight line. This property forms the external occlusions in real LiDAR-captured data. To simulate the occlusion naturally, we first transform the point cloud of 3D scene and the point cloud of inserted human into the range image I^Q and I^q, respectively. Then, we merge two range images based on the distance D to LiDAR to insert a synthetic human instance into the scene. the formula is

$$\begin{aligned} I^{Q'} = &\{I_{ij}^Q | D(I_{ij}^Q) < D(I_{ij}^q)\} \cup \\ &\{I_{ij}^q | D(I_{ij}^Q) \geq D(I_{ij}^q)\}, \end{aligned} \tag{1}$$



Stage 1: Instance-to-Scene Representation Transfer *Stage 2: Synthetic-to-Real Feature Alignment* *Stage 3: Fine-grained Perception enhancement*

Figure 2. **Pipeline of our method.** "PC", "RI", "RC" stand for Point Cloud, Range Image, Receptive Control, respectively. The individuals painted with yellow represent real humans, while those with pink represent synthetic humans. For stage1, we introduce range image bridged insertion, a module insert parametric model into existing dataset to create our natural synthetic data. We train our detector on the data to produce initial pseudo-labels. In stage 2, we employ unsupervised bi-directional filter to improve the quality of pseudo-label. Then, Synthetic-to-real feature alignment is applied to enhance the generalize ability of our detector to real human. During stage 3, we utilize human structural knowledge to boost the performance of the model. Finally, based on the obtained high-quality pseudo-labels, fine-tuning is used to make the model totally converge to identify real humans.

where $0 < i < H$ and $0 < j < W$. Finally, we transform $I^{Q'}$ back into point cloud, and fit bounding box to the synthetic human instance as our synthetic data label. For each selected scene, we repeat this process multiple times to create more synthetic human in the same scene. To prevent synthetic human collide into each other, we employ Intersection over Union (IoU) threshold and bounding box center distance threshold to filter out invalid insertion.

Mask-constrained receptive field control. By using the simulated scene with "annotated" human instances, we can train our model directly. However, as we only "annotated" synthetic humans, those real humans without annotations will be considered as irrelevant backgrounds, inevitably leading the model to ignore real human instances. To address this, we generate a mask to constrain the receptive field to areas where synthetic humans locate while real humans do not. We choose the renowned anchor-free 3D detection network CenterPoint [44] as our backbone, which gives BEV heatmap and bounding box as outputs, indicating the human's location and rotation. To generate mask M for vacant ground without objects first, we voxelize the scene and calculate the number of empty voxels above each BEV grid. If the number exceeds a threshold, we regard it as the vacant ground. During the training phase, this mask M undergoes logical *or* operation with the ground truth heatmap y created by the synthetic label to generate M^*, which represents the ground without real humans. Centerpoint [44] utilize Gaussian focal loss for heatmap loss and ℓ_2-norm for hounding box loss. We modify the heatmap loss L_{hm} to work with our receptive control mask:

$$L_{hm} = -\sum_{i=1}^{H'}\sum_{j=1}^{W'} \begin{cases} 0 & \text{if } M^*_{ij} = 0 \\ (1-x_{ij})_1^{\beta}\ln(x_{ij}+\varepsilon) & \text{if } M^*_{ij}, y_{ij} = 1 \\ x_1^{\beta}(1-y_{ij})_2^{\beta}\ln(1-x_{ij}+\varepsilon) & \text{otherwise,} \end{cases} \tag{2}$$

where x represents the predict heatmap, H',W' represents the shape of heatmap, β_1 and β_2 control the contribution of each grid, and ε prevents $\ln()$ to take on a very small value. The final loss L of our 3D detector is defined as:

$$L = L_{hm} + L_{bbox}. \tag{3}$$

Note that the mask is only used during training procedure. With "annotated" data and mask-constrained receptive field control, we trained our detector using only the generated synthetic data. We then infer purely on training split to generate pseudo-labels of real humans for the next stage.

3.2. Synthetic-to-Real Feature Alignment

While our first stage effectively generates synthetic data with a high degree of naturalness, a domain gap exists between synthetic and real humans in aspects such as clothing, pose distribution and overall shape. We utilize BEV feature alignment to narrow the gap in order to augment the detector's ability to detect real humans. To reduce false alignment, we first utilize a bi-directional filter to select high-quality pseudo-labels for feature alignment.

Bi-directional filter. Drawing inspiration from [46], we leverage a bi-directional multi-object tracking algorithm to examine the temporal consistency of pseudo-labels to filter out erroneous ones. Typically, an unsupervised tracking algorithm contains two parts: one to predict the movement of tracklets and the other to match the prediction with a tracklet. We employ 3D Kalman filter [37] to predict the movement of the human and greedy algorithm to match detection to each tracklet. After matching, the Kalman filter parameter is updated with the matched detection data. All tracklets are classified into either long or short tracklet based on their temporal length. We discard short tracklet due to their low temporal consistency. In stationary view cases, we check whether the object has relocated for a long tracklet additionally to filter out errors caused by background. To improve the quality of the tracking, we simply run the tracking in both temporal directions and merge the results based on IoU and bounding box centre distance.

BEV feature alignment. With refined pseudo-labels, we can perform BEV feature alignment to enhance our detector's ability to detect real humans. We align feature through loss L_{S_2R}, which contains two parts: $L_{f_2\bar{f}}$ and L_{norm}. $L_{f_2\bar{f}}$ diminish the gap between synthetic and real human features. L_{norm} bounds the activate and prevents features from collapsing into a zero vector, thereby averting the occurrence of "dead neurons" in the neural network. We first gather synthetic human feature F_s and real human feature F_r from the BEV feature. Afterwards, we calculate the means of simulated and real human features, denoted as $\overline{F_s}$ and $\overline{F_r}$. we present the formation for L_{S_2R} below:

$$
\begin{aligned}
&L_{s_2r} = \left(\overline{F_s} - \overline{F_r}\right)^2, \\
&L_{norm} = \frac{1}{|F_s|} \sum_{f_s \in F_s} R\left(|1 - \|f_s\|_2| - \Delta var\right)^2 \\
&\qquad + \frac{1}{|F_r|} \sum_{f_r \in F_r} R\left(|1 - \|f_r\|_2| - \Delta var\right)^2, \\
&L_{S2R} = \beta_3 * L_{s_2r} + \beta_4 * L_{\text{norm}},
\end{aligned}
\tag{4}
$$

where $R(\cdot)$ denotes Rectified Linear Unit (ReLU) and $\|\cdot\|_2$ denotes ℓ_2-norm. Δvar is introduced to regulate the maximum allowable distance. β_3, β_4 are constant weights for loss. With cleaned pseudo-label provided by our bidirectional multi-object filter and BEV feature alignment guiding the model, we retrain the model on our "annotated" scene. After that, the model is inferred on training split again to generate higher-quality pseudo-labels.

3.3. Fine-Grained Perception Enhancement

After BEV feature alignment, the global features of real humans become close to these of synthetic humans, and we can further transfer more fine-grained knowledge of synthetic data to enhance the capability of model to identify detailed human-specific features. In this stage, we improve the detection performance by transferring human structural semantics to the model. We also update the receptive field based on previous stage's pseudo-labels.

Human structural knowledge transfer. Human bodies comprise various parts, such as arms, legs, and trunks. This knowledge is crucial in identifying humans, especially when occlusion occurs, because the presence of these body parts helps differentiate humans from objects. Based on this, we propose to transfer the knowledge of body parts to further enhance the model's feature extraction ability. As shown in Fig. 2, we select six key joints representing the arms, legs, trunk, and head. We consider these to be the most indicative of human structure. Since the realistic occlusion in our synthetic data may make some body parts invisible, we filter out invisible parts based on the number of points within that part away from the closest key joint. Finally, we add joint prediction task heads to our backbone, parallel to the original task head that outputs the heatmap and bounding box. These task heads share the same loss function 2 mentioned before.

Receptive field update. Leveraging the detection results obtained from the instance-to-scene representation transfer and synthetic-to-real feature alignment stage, we can now identify a substantial proportion of real humans. We utilize these two stages' results to update our receptive field so that our model can interact with more areas in the scene. Specifically, we project the bounding box of a pseudo-label to BEV and expand the box (e.g., 2m×2m) to form the mask. This mask is then inverted and combined with the original mask M using a logical *or* operation, creating a larger mask M'. This expanded mask contains more areas of the scene, which can improve the robustness of the model, thereby enhancing its performance.

3.4. Fine-tuning

After previous three stages, we have obtained relative high-quality pseudo-labels. To make the detector further converge to identify features of real humans, we fine-tune the model based on raw point cloud data and pseudo-label supervision and obtain the final results.

4. Experiments

In this section, we first provide the detailed experimental setup, then evaluate the proposed method on two human-centric datasets, *i.e.*, HuCenLife and STCrowd. Next, extensive ablation studies are presented to analyze the effectiveness of each key component in our method. Subsequently, we demonstrate how performance is affected by varying amounts of synthetic data. Finally, we show our feature extractor's potency by fine-tuning with various amount of ground truth data.

4.1. Experimental setup

Datasets. We evaluate our approach on two datasets: HuCenLife [41] and STCrowd [10]. To the best of our knowledge, only these two datasets are human-centric so far. HuCenLife features rich human-environment interaction and abundant human pose. It contains 212 sequences, where

Table 1. **3D Human detection performance on HuCenLife dataset**. "$*$" shows the result acquired by full supervision.

	AP(0.25)	AP(0.50)	AP(1.0)	Prec(0.25)	Prec(0.5)	Prec(1.0)	Recall(0.25)	Recall(0.5)	Recall(1.0)	mPrec	mRecall	mAP
CenterPoint*[44]	59.24	73.18	74.65	59.65	70.27	71.61	66.46	78.29	79.79	67.18	74.84	69.02
DBSCAN[13]	-	-	-	7.95	10.88	13.71	15.16	20.76	26.15	10.85	20.69	-
MODEST[45]	0.98	40.65	57.63	8.34	**48.02**	**58.70**	9.58	55.21	67.50	38.35	44.10	33.09
OYSTER[46]	24.57	36.33	39.10	28.30	35.76	37.86	45.35	57.29	60.66	33.97	54.44	33.33
Ours	**55.20**	**64.90**	**66.36**	**39.49**	44.50	45.22	**70.85**	**79.84**	**81.13**	**43.07**	**77.27**	**62.15**

Table 2. **3D Human detection performance on STCrowd dataset**. "$*$" shows the result acquired by full supervision.

	AP(0.25)	AP(0.50)	AP(1.0)	Prec(0.25)	Prec(0.5)	Prec(1.0)	Recall(0.25)	Recall(0.5)	Recall(1.0)	mPrec	mRecall	mAP
CenterPoint*[44]	80.86	86.87	87.80	61.8	64.94	65.40	88.90	93.41	94.07	64.05	92.13	85.17
DBSCAN[13]	-	-	-	6.33	10.91	14.34	5.70	9.82	12.90	10.53	9.47	-
MODEST[45]	0.01	15.07	38.58	0.65	35.75	**64.16**	0.45	24.40	43.80	33.52	22.88	17.89
OYSTER[46]	16.84	23.85	25.03	32.68	39.61	40.59	40.58	49.18	50.39	37.63	46.72	21.90
Ours	**58.38**	**70.94**	**72.28**	**41.78**	**47.21**	47.95	**72.31**	**81.71**	**82.99**	**45.64**	**79.00**	**67.20**

3931 frames are used for train and 2254 frames are used for validation. It uses 128-beam LiDAR with $45°$ vertical front of view (FOV). STCrowd emphasizes on pedestrian, which features dense human crowd. It contains 84 sequence, where 5263 frames are used for train and 2992 frames are used for validation. It uses 128-beam LiDAR with $90°$ vertical FOV. Note that no ground truth are used during training for both datasets.

Implementation details. We generate synthetic human data using SMPL [19] model with pose and shape parameter from SURREAL [34]. Specifically, we first randomly choose a sequence, then randomly choose a frame within this sequence. After that we put random number of synthetic humans on ground as 3.1 described. For HuCenLife we set the detection range to [25.6, 51.2] meters for the X and Y axis and [-2.5, 7.5] meters for the Z axis. For STCrowd, we focus on the cropped point cloud range with [30.72, 40.96] meters for the X and Y axis and [-4, 1] for the Z axis. Moreover, to fulfill our fine-grained detection necessity, we use denser voxel than the offical setting of STCrowd. Specifically, we use [0.025, 0.05, 0.25] for HuCenLife and [0.03, 0.04, 0.125] for STCrowd. For post-processing of detection results, we use a circle NMS method with a threshold of 0.2 for all the experiments.

Metrics. Following HuCenLife [41] and STCrowd [10] official metrics, we use Precision, Recall and Average Precision (AP) with 3D center distance threshold $D = \{0.25, 0.5, 1\}$. Furthermore, we compute the mean Average Precision(mAP), mean Precision (mPrec), mean Recall (mRecall) by averaging AP, Precison and Recall.

Baselines. We conduct experiments on the following three unsupervised detection baselines. **DBSCAN** [13] performs density-based clustering on point cloud, thus generates class-agnostic pseudo labels. There is no training required. **MODEST** [40] first calculates the persistence point score (PP score) by identifying ephemeral points in repeated traversals, and then uses these scores to conduct DBSCAN clustering. Finally, they train the detector using pseudo labels generated after clustering for 10 rounds, where at each round MODEST utilizes PP score to refine the output . This methodology has been optimized for peak performance on HuCenLife and STCrowd dataset based on official code. **OYSTER** [46] directly conducts DBSCAN clustering algorithm to generate initial pseudo labels for training. After each round of training, OYSTER conduct an unsupervised bi-directional tracking to filter temporally inconsistent objects. To ensure optimal performance, we replicated the code, following the methodologies detailed in the paper and optimizing for the best, such as we discard its pseudo-label refinement. The refinement, which assigns a uniform size to all bounding boxes in the same tracklet, is effective for rigid bodies, but not for instances as flexible as human. Note that for our method, MODEST, and OYSTER, we choose labels coming from the best instead of the last epoch as the next round's pseudo label for fair comparison, which reduce uncertainty caused by selection. We use the same CenterPoint as backbone for our method, MODEST, and OYSTER.

4.2. Results

Results on HuCenLife. We provide comparison against SOTA methods on HuCenLife in Tab. 1. **DBSCAN** performs poorly since it only considers local density rather than structural nor semantic information, thus generates a large amount of objects instead of centering on human. Since it doesn't provide confidence score, we only report its precious and recall. **MODEST** leverages PP score for clustering, which introduces motion information to DBSCAN, thus achieving better performance. However, MODEST uses the same PP score to filter each training round's output, which may weaken the network's fine-grained detection capacity. **OYSTER** gains only a marginal improvement in mAP compared to MODEST, despite its better performance in traffic scenario [46]. This disparity can be attributed to DBSCAN's poor performance in human-centric 3D detection, which is OYSTER's initial pseudo-label generation method.

Our method surpasses MODEST by 87.8% and OYSTER by 86.5% in terms of mAP. Initially, our approach involves transferring knowledge from synthetic human instances to real-world scenes. Subsequently, we apply synthetic-to-real feature to generalize models detection ability. In the final stage, we utilize body skeleton as an ad-

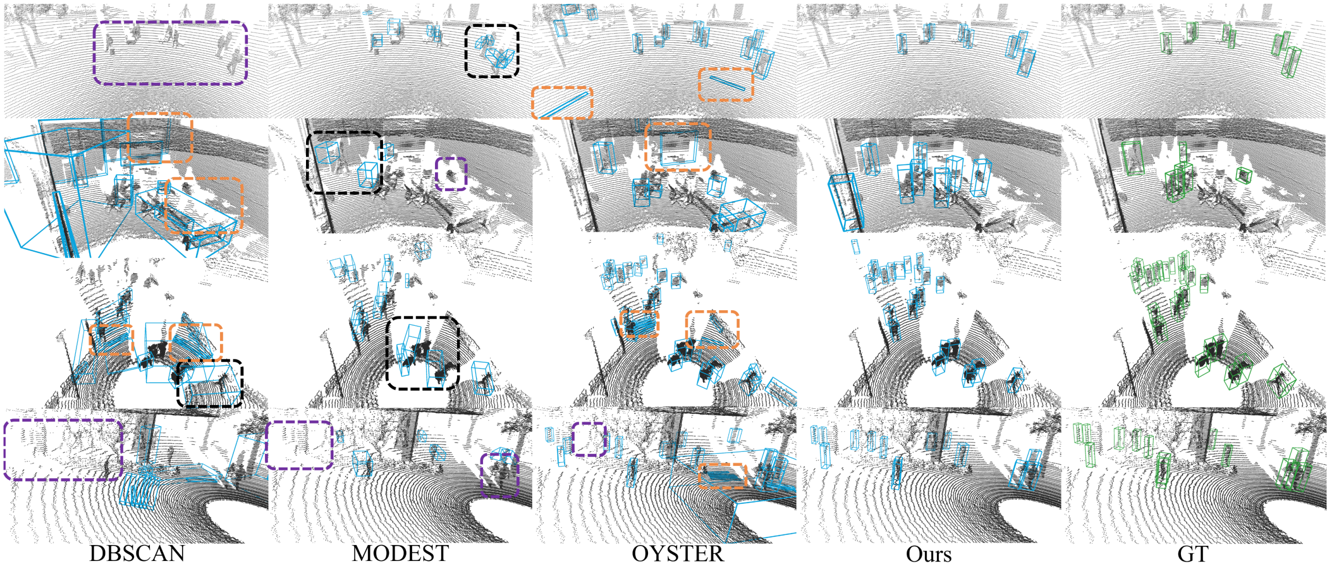


Figure 3. **Detection visualization.** The first and second rows demonstrate results on HuCenLife[41]. The third and forth rows show results on STCrowd[10].

Table 3. **Ablation study.** We conduct all following experiments on HuCenLife dataset. "mPrec" stands for mean precision."mRecall" stands for mean recall. "ID" stands for experiment ID. "RIB insertion" stands for range image based insertion. "RF control" stands for receptive field control. "S2R align" stand for BEV feature alignment. "RF update" stands for receptive field update. "HSK transfer" stands for human structural knowledge transfer. For E_0, we use DBSCAN to generate pseudo-label for training. For E_1-E_7, we use the same 50k frames of generated synthetic human data for fair comperasion.

	Instance-to-scene representation transfer		Synthetic-to-real feature alignment		Fine-grained perception enhancement			Fine-tuning	Performance					
ID	RIB insertion	RF control	Fine-tuning	S2R align	Retrain	RF update	HSK transfer	Retrain	AP(0.25)	AP(0.5)	AP(1.0)	mPrec	mRecall	mAP
E_0									14.49	23.09	26.33	18.97	44.97	21.31
E_1	✓								22.69	30.28	32.10	**51.82**	40.73	28.36
E_2	✓	✓							32.03	42.82	45.25	28.15	68.21	40.03
E_3	✓	✓	✓						27.58	38.49	41.29	21.29	69.42	35.78
E_4	✓	✓	✓	✓					32.07	44.19	46.57	37.76	62.12	40.94
E_5	✓	✓	✓	✓	✓				46.32	58.52	60.55	14.54	80.95	55.13
E_6	✓	✓	✓	✓	✓	✓			48.07	61.63	63.43	16.54	**82.41**	57.71
E_7	✓	✓	✓	✓	✓	✓	✓		52.44	61.56	63.08	24.88	79.26	59.02
E_8	✓	✓	✓	✓	✓	✓	✓	✓	**55.20**	**64.90**	**66.36**	43.07	77.27	**62.15**

ditional supervisory. Notably, our method does not rely on clustering algorithms, but benefits from high-quality synthetic data from beginning. Furthermore, our modules is specifically designed to guide towards both generalization and meticulous detail, which contribute to its notable performance improvements.

Results on STCrowd. In Tab. 2, We provide comparison against SOTA methods on STCrowd. We observe that **DBSCAN** perform worse, because STCrowd is more dense, with more crowded scenarios, making DBSCAN tend to cluster multiple individuals who are density-reachable into a singe cluster. Apart from that, we notice that performance for **OYSTER** and **MODEST** all degrades, with MODEST experiencing a more pronounced drop. MODEST relys on PP score calculated by points distance between traversals for initial pseudo-labels generation and refinement. However it is hard to match points between frame in dense crowd common in STCrowd. By comparison, our method still performs closely to fully supervised method (78.9% in term of mAP) and dominates the table, demonstrating our approach's ability in accurately locating each individual within dense crowds.

Qualitative results. We provide qualitative results on both HuCenLife and STCrowd in Fig. 3. For **OYSTER** and **DBSCAN**, they suffer a lot from false positive detection (shown as **orange** boxes). Unlike our approach, they are class-agnostic, which consider all the objects after clustering as human, such as walls, television, and even ground that failed to be removed from ground removal process. For **MODEST**, it relys on motion detector to produce and filter pseudo-labels. In human-centric detection, the challenge arises from the relatively small amplitude of human movements and the possibility that only a portion of the person's body is in motion. As a result, motion detector provides bounding box with error in size and position in these scenario (shown as **black** boxes). Furthermore, **OYSTER**, **MODEST**, and **DBSCAN** all suffers from false-negative detection, for the same reason (shown as **purple** boxes).

Table 4. **Performance on synthetic frame amount**. We report performance on different amount of generated frames.

Synthetic frames amount	AP(0.25)	AP(0.50)	AP(1.0)	mAP
5k	46.10	56.16	58.82	53.70
10k	53.50	62.70	64.28	60.16
25k	52.35	62.44	63.81	59.53
50k	55.20	64.90	66.36	62.15

Despite our method generating false positive detection on the second row, it still outperforms other methods. This truly demonstrates the effectiveness of our approach.

4.3. Ablation study

We present our ablation study in Tab. 3. In this part we will analyze our method's effectiveness module by module.

Effect of instance-to-scene representation transfer. Range image bridged insertion brings spatial property of LiDAR to mesh-formed synthetic humans. As we can see in $E_0 \rightarrow E_1$, with the same training configuration, our synthetic pseudo labels outperform the class-agnostic DBSCAN with around 30% improvement. Apart from that, we can observe that receptive field control reduces the erroneous instruction towards the model, which improves the performance ($E_1 \rightarrow E_2$).

Effect of synthetic-to-real feature alignment. E_3 demonstrates that, in the absence of additional refinement, training with pseudo-labels is impractical. This ineffectiveness arises due to the inferior quality of pseudo-labels compared to synthetic labels, even after tracking filtering. The introduction of BEV feature alignment ($E_3 \rightarrow E_4$) serves as a corrective measure. This alignment not only bridges the gap between pseudo and synthetic labels but also enhances the model's capacity to accurately detect real humans, thereby significantly improving overall performance.

Effect of Fine-grained perception enhancement. The receptive field update expands the models perception capacity, which allows it to interact with more background objects. This improves mAP ($E_5 \rightarrow E_6$) by 4.7%. Utilizing human structural knowledge ($E_6 \rightarrow E_7$), the model further improves its feature extraction ability, gains a improvement of 2.2% in mAP. $E_7 \rightarrow E_8$ shows fine-tuning on original training data. This gives the model 5.3% improvement.

Effect of different amount of synthetic data. In this section, we explore the impact of varying amounts of synthetic data on our algorithm. We report the results in Tab. 4. The table suggests that as the number of synthetic frames increases, there is an improvement in performance across the evaluated metrics, showing the positive impact of varying amounts of synthetic data on the algorithm's effectiveness. In detail, even if we use just 5k synthetic frames (about 27% more than the original annotated training data) for training, the performance still surpass current SOTA methods over 60%. This further demonstrate that our method has a better capacity for capturing human's both surface level and fine-grained representation. Furthermore, we observed enhancement over all metrics from 5k to 10k, whereas the incremental gains become marginal when we use around 50k synthetic frames. This observation demonstrates that increasing the number of synthetic frames improve performance but the payback start to diminish as frame number increase. Hence, we selected 50k frames to strike the balance between performance and efficiency for training.

Table 5. **Performance on fine-tuning with GT**. We report performance on fine-tuning with small amount of ground truth data. *GT* column stands for how much ground truth data is used. "$*$" presents the performance of full supervision using the same proportion of ground truth data.

GT	AP(0.25)	AP(0.5)	AP(1.0)	mAP
0%	55.20	64.90	66.36	62.15
1%*	15.40	35.61	43.00	31.34
1%	**55.92**	**70.22**	**71.95**	**66.03**
10%*	29.02	54.85	61.22	48.36
10%	**63.01**	**73.63**	**74.97**	**70.54**
20%*	55.35	68.68	70.79	64.94
20%	**63.54**	**74.78**	**76.29**	**71.53**
100%*	59.24	73.18	74.65	69.02
100%	**64.25**	**75.66**	**77.02**	**72.31**

4.4. Effect of feature extractor

We evaluate our feature extractor's effectiveness in Tab. 5 by fine-tuning with varying proportions of ground truth data. Note that we freeze all the parameters and weights for the feature extractor. Remarkably, our feature extractor's performance exceeds that of the fully supervised approach by a substantial margin, when utilizing an equivalent amount of ground truth data. With only 10% of ground truth data employed for fine-tuning, our feature extractor surpasses the performance of the fully supervised method trained on 100% ground truth data. Furthermore, there is a notable 4.7% improvement over the fully supervised method when utilizing the full 100% ground truth data for fine-tuning. This noteworthy result further underscores the efficacy of our feature extractor, showcasing its ability to adapt and refine feature representations.

5. Conclusions

We introduce an unsupervised 3D detection approach for human-centric scenarios, which transfers knowledge from synthetic human models to actual scenes. We develop effective modules for instance-to-scene representation transfer and synthetic-to-real feature alignment to overcome the disparities in data representations and feature distributions between synthetic models and real-world point clouds. Impressively, our method outperforms existing state-of-the-art methods with a significant margin, and nearly matching the results of fully supervised methods.

References

[1] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 1, 2

[2] Igor Bogoslavskyi and C. Stachniss. Fast range image-based segmentation of sparse 3d laser scans for online operation. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 163–169, 2016. 2

[3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2

[4] Runnan Chen, Xinge Zhu, Nenglun Chen, Dawei Wang, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Towards 3d scene understanding by referring synthetic models. *arXiv preprint arXiv:2203.10546*, 2022. 3

[5] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, ZHU Xinge, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[6] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[7] Runnan Chen, Xinge Zhu, Nenglun Chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. Bridging language and geometric primitives for zero-shot point cloud segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5380–5388, 2023. 3

[8] Runnan Chen, Xinge Zhu, Nenglun Chen, Dawei Wang, Wei Li, Yuexin Ma, Ruigang Yang, Tongliang Liu, and Wenping Wang. Model2scene: Learning 3d scene representation via contrastive language-cad models pre-training. *arXiv preprint arXiv:2309.16956*, 2023. 3

[9] Peishan Cong, Xinge Zhu, and Yuexin Ma. Input-output balanced framework for long-tailed lidar semantic segmentation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3

[10] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19576–19585, 2022. 1, 2, 5, 6, 7

[11] Peishan Cong, Yiteng Xu, Yiming Ren, Juze Zhang, Lan Xu, Jingya Wang, Jingyi Yu, and Yuexin Ma. Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 461–469, 2023. 3

[12] 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. 2

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, 1996. 3, 6

[14] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[15] Aleksey Golovinskiy and Thomas A. Funkhouser. Min-cut based segmentation of point clouds. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 39–46, 2009. 2

[16] 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. 2

[17] Jialian Li and etc. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *CVPR*, pages 20502–20512, 2022. 2

[18] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347*, 2023. 3

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2015. 6

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[21] Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, and Yuexin Ma. See more and know more: Zero-shot point cloud segmentation via multi-modal visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21674–21684, 2023. 3

[22] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, pages 6120–6127, 2019. 2

[23] Mahyar Najibi, Jingwei Ji, Yin Zhou, C. Qi, Xinchen Yan, Scott M. Ettinger, and Drago Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *European Conference on Computer Vision*, 2022. 2

[24] Mahyar Najibi, Jingwei Ji, Yin Zhou, C. Qi, Xinchen Yan, Scott M. Ettinger, and Drago Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8568–8578, 2023. 3

[25] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al.

Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 3

[26] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Tai Wang, Xinge Zhu, and Yuexin Ma. Sam-guided unsupervised domain adaptation for 3d segmentation. *arXiv preprint arXiv:2310.08820*, 2023. 3

[27] Xidong Peng, Xinge Zhu, and Yuexin Ma. Cl3d: Unsupervised domain adaptation for cross-lidar 3d detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2047–2055, 2023. 3

[28] Jordi Pont-Tuset, Pablo Arbeláez, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:128–140, 2017. 2

[29] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3283–3292, 2021. 3

[30] 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. 2

[31] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997. 2

[32] 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

[33] Haofei Tian, Yuntao Chen, Jifeng Dai, Zhaoxiang Zhang, and Xizhou Zhu. Unsupervised object detection with lidar clues. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5958–5968, 2020. 2

[34] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6

[35] Yangtao Wang, XI Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14523–14533, 2022. 2

[36] Yu-Quan Wang, Yuntao Chen, and Zhaoxiang Zhang. 4d unsupervised object discovery. *ArXiv*, abs/2210.04801, 2022. 2

[37] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020. 5

[38] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4376–4382. IEEE, 2019. 2

[39] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 3

[40] Jenny Xu and Steven L Waslander. Hypermodest: Self-supervised 3d object detection with confidence score filtering. *arXiv preprint arXiv:2304.14446*, 2023. 3, 6

[41] Yiteng Xu, Peishan Cong, Yichen Yao, Runnan Chen, Yuenan Hou, Xinge Zhu, Xuming He, Jingyi Yu, and Yuexin Ma. Human-centric scene understanding for 3d large-scale scenarios. *ArXiv*, abs/2307.14392, 2023. 1, 2, 5, 6, 7

[42] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 2

[43] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposal-contrast: Unsupervised pre-training for lidar-based 3d object detection. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022. 3

[44] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 2, 4, 6

[45] Yurong You, Katie Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Learning to detect mobile objects from lidar scans without labels. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1120–1130, 2022. 2, 3, 6

[46] Lunjun Zhang, Anqi Joyce Yang, Yuwen Xiong, Sergio Casas, Bin Yang, Mengye Ren, and Raquel Urtasun. Towards unsupervised object detection from lidar point clouds. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9317–9328, 2023. 2, 3, 4, 6

[47] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17619–17629, 2023. 2, 3

[48] 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. 2

[49] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *ECCV*, pages 581–597. Springer, 2020. 1

[50] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6807–6822, 2021. 1, 2