EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI

Tai Wang[1*], Xiaohan Mao[1,2*], Chenming Zhu[1,3*], Runsen Xu[1,4], Ruiyuan Lyu[1,5], Peisen Li[1,5],
Xiao Chen[1,4], Wenwei Zhang[1], Kai Chen[1], Tianfan Xue[4], Xihui Liu[3], Cewu Lu[2],
Dahua Lin[1,4], Jiangmiao Pang[1✉]

[1]Shanghai AI Laboratory [2]Shanghai Jiao Tong University [3]The University of Hong Kong
[4]The Chinese University of Hong Kong [5]Tsinghua University
[*]Equal contribution [✉]Corresponding author

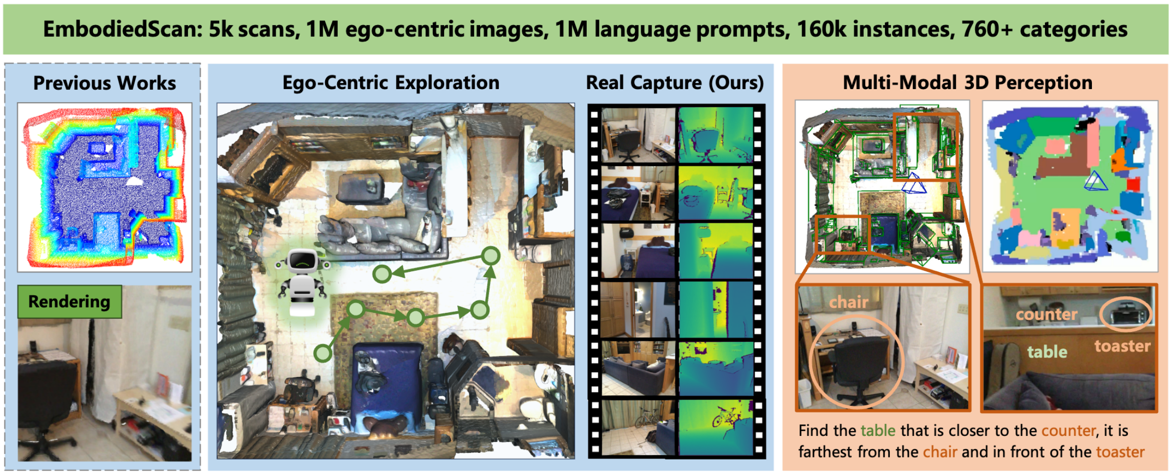

EmbodiedScan: 5k scans, 1M ego-centric images, 1M language prompts, 160k instances, 760+ categories
Previous Works
Rendering
Ego-Centric Exploration
Real Capture (Ours)
Multi-Modal 3D Perception
chair
counter
toaster
table
Find the table that is closer to the counter, it is farthest from the chair and in front of the toaster

Figure 1. EmbodiedScan provides a multi-modal, ego-centric 3D perception dataset with massive real-scanned data and rich annotations for indoor scenes. It benchmarks language-grounded holistic 3D scene understanding capabilities for real-world embodied agents.

Abstract

In the realm of computer vision and robotics, embodied agents are expected to explore their environment and carry out human instructions. This necessitates the ability to fully understand 3D scenes given their first-person observations and contextualize them into language for interaction. However, traditional research focuses more on scene-level input and output setups from a global view. To address the gap, we introduce EmbodiedScan, a multi-modal, ego-centric 3D perception dataset and benchmark for holistic 3D scene understanding. It encompasses over 5k scans encapsulating 1M ego-centric RGB-D views, 1M language prompts, 160k 3D-oriented boxes spanning over 760 categories, some of which partially align with LVIS, and dense semantic occupancy with 80 common categories. Building upon this database, we introduce a baseline framework named Embodied Perceptron. It is capable of processing an arbitrary number of multi-modal inputs and demonstrates remarkable 3D perception capabilities, both within the two series of benchmarks we set up, i.e., *fundamental 3D perception tasks and language-grounded tasks, and in the wild.*

1. Introduction

Consider an embodied agent operating in an indoor environment. It commences its journey devoid of any prior knowledge about the scene, guided only by an initial instruction. As it begins to explore, it recognizes objects in context and acts with goals along with language interaction. In this process, a commonly needed, fundamental perception capability is to establish a *holistic* 3D scene understanding given *ego-centric* observations. This understanding operates at the scene level, covers both object semantics and scene geometry, and can be grounded in language descriptions.

Nonetheless, subtle but significant discrepancies exist between this expectation and research problems examined within the computer vision community. Most previous studies have primarily revolved around scene-level input and output problems from a global view [13, 34, 40], *i.e.*, taking reconstructed 3D point clouds or meshes as inputs and predicting 3D object bounding boxes or segmenting point clouds. Regarding data, earlier datasets targeting ego-centric RGB-D inputs are either too small [12, 45] or lack comprehensive annotations [6, 51] to support the aforemen-

Dataset	#Scans	#Imgs	#Objs	#Cats	#Prompts	Ego Capture	3D Annotations
Replica [46]	35	-	-	-	-	✗	✗
NYU v2 [12]	464	1.4k	35k	14	-	✓	✗
SUN RGB-D [45]	-	10k	-	37	-	Mono.	Box
ScanNet [13, 39]	1513	264k	36k	18	52k [8]	✓	Seg., Lang.
Matterport3D [6]	2056	194k	51k	40	-	Multi-View	Seg.
3RScan [51]	1482	363k	-	-	-	✓	Seg.
ArkitScenes [2]	5047	450k	51k	17	-	✓	Box
HyperSim [38]	461	77k	-	40+	-	Mono. & Syn.	Box
EmbodiedScan	5185	890k	160k	762	970k	✓	Box, Occ., Lang.

Table 1. Comparison with other 3D indoor scene datasets. "Cats" refers to the categories with box annotations for the 3D detection benchmark. EmbodiedScan features more than 10× categories, prompts, and the most diverse annotations. The numbers are still scaling up with further annotations. Mono./Syn./Lang. means Monocular/Synthetic/Language.

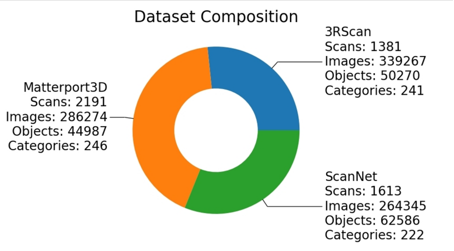


Figure 2. Dataset composition. EmbodiedScan is composed of three data sources and has similar scans, images, objects, and categories in each of them.

tioned research. It is also not feasible to generate such realistic views by rendering from the existing imperfect meshes. On the other hand, since we cannot trivially obtain the reconstruction of a new environment, models trained with scene-level input are not directly applicable in practice.

To bridge this divide, we introduce a multi-modal, ego-centric 3D perception dataset and benchmark for holistic 3D scene understanding, termed *EmbodiedScan*, aimed at facilitating real-world embodied AI applications (Fig. 1). This dataset exploits existing large-scale 3D scene datasets [6, 13, 51] but re-purposes them for continuous scene-level perception from the first-view RGB-D streams. Unlike previous works that offer only point segmentation labels with limited semantics, we employ a SAM-assisted [22] pipeline to annotate objects with oriented 3D bounding boxes and generate language prompts on top. Consequently, EmbodiedScan provides more than 5k scans, nearly 1M ego-centric RGB-D images, and multi-modality annotations, covering 3D oriented boxes with more than 160k instances spanning over 760 categories, dense semantic occupancy with 80 common categories, and 1M language descriptions focusing on spatial relationships among objects.

Built upon this dataset, we devise a baseline framework for ego-centric 3D perception, *Embodied Perceptron*. It accepts RGB-D sequences and texts as inputs and manifests scalability to any number of views input with encoders shared across different tasks. With the encoded 2D and 3D features, we employ dense fusion and isomorphic multi-level fusion across them guided by the perspective projection to produce 3D scene and object representations, which are further processed to decode occupancy and 3D box predictions. The derived 3D representations can be further integrated with text embeddings for 3D visual grounding, thus supporting language-grounded applications.

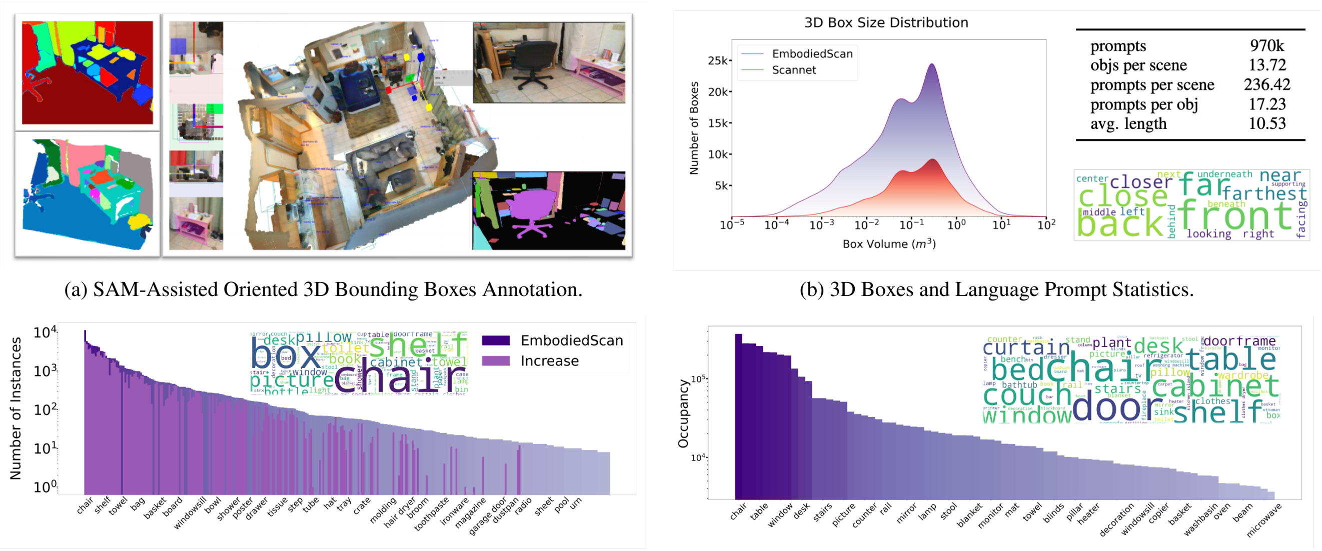
We establish two series of benchmarks on EmbodiedScan: 1) fundamental 3D perception benchmarks focusing on traditional tasks, including 3D detection and semantic occupancy prediction under different input settings, and 2) a language-grounded scene understanding benchmark with 3D visual grounding as a preliminary exploration. Experimental results validate the effectiveness of our baseline model on EmbodiedScan and demonstrate its generalization ability in the wild. Detailed analysis further underscores the value of EmbodiedScan and highlights the primary challenges posed by this new setup.

2. Related Work

3D Scene Datasets. The development of 3D scene understanding has benefited from large-scale, high-quality datasets like KITTI [16] and SUN RGB-D [45]. These foundational datasets have paved the way for subsequent larger and more diverse collections targeting indoor [6, 13, 38, 51] and driving scenes [4, 7, 32, 47]. However, compared to autonomous driving datasets, those meant for indoor scenes still lack variety in terms of scenes and object diversity (Tab. 1). In contrast, EmbodiedScan provides a large amount of multi-modal data with much richer annotations. Furthermore, it differs by placing an emphasis on the ego-centric perspective within its setup, a feature often overlooked in previous works [2, 13].

Except for these conventional dataset works, Omni3D [3] integrates urban and indoor datasets for monocular 3D detection. Our focus, however, lies in indoor scenes due to their unique challenges but has a larger amount of data and annotations, *e.g.*, more than 3× images and categories with more than 10 instances. In addition, we offer a comprehensive exploration of more general problems for ego-centric 3D perception, such as continuous perception and visual grounding. Other embodied AI datasets like HM3D [36, 56] and HSSD [21] provide ample interaction opportunities but can suffer from poor transferability to real-world scenarios due to their imperfect meshes or synthetic data. Conversely, EmbodiedScan is based on real-scanned RGB-D images, offering a more realistic playground for model training.

3D Object Detection & Occupancy Prediction. 3D detection and occupancy prediction, as fundamental tasks in 3D perception, focus on different aspects of 3D scene understanding. The former focuses on recognizing foreground objects through a sparse and efficient representation - a set of 3D cuboids corresponding to instances of interest -



(a) SAM-Assisted Oriented 3D Bounding Boxes Annotation.

(b) 3D Boxes and Language Prompt Statistics.

(c) Instance Statistics (Increase w.r.t. ScanNet).

(d) Occupancy Statistics.

Figure 3. EmbodiedScan annotation and statistics. (a) UI for 3D box annotation. We select keyframes and generate their SAM masks with corresponding axis-aligned boxes. With simple clicks, annotators can create 3D boxes for target objects and further adjust them with reference in three orthogonal views and images. (b) Small boxes ($< 1m^3$) increase more & prompt statistics. objs/avg./des. refer to objects/average/descriptions. (c) We show the number of instances per category (300 classes). For categories that exist in ScanNet, we plot the absolute increase and observe a significant improvement. (d) We plot the occupancy distribution for each category and see a different word cloud distribution. These two clouds show different aspects, occupied space vs. number of instances, of this dataset.

while the latter offers a dense, structured pattern that benefits downstream planning. The research community has developed solutions ranging from single-modality, such as LiDAR-based [23, 29, 34, 40, 42, 57, 59] or camera-only approaches [28, 41, 44, 52–54], to multi-modality techniques [26, 30, 35, 50, 58]. Recently, occupancy as a representation, due to its potential in handling unknown semantics and irregular object shapes, has gained more attention [5, 19, 25, 43, 49, 55]. Given their distinctive focuses, we selected these two tasks to form the fundamental 3D perception track on EmbodiedScan.

Previous works on indoor scenes mainly centered around 3D detection with limitations in object orientations, semantic categories, and input format [29, 34, 40]. In practice, a model is expected to perceive the environment during egocentric exploration, ultimately providing a holistic understanding inclusive of rich semantics, scene geometry, and object poses. To this end, EmbodiedScan and our proposed framework, Embodied Perceptron, provide this necessary data foundation and baseline methodology.

Language-Grounded 3D Scene Understanding. Language plays a crucial role in human-computer interaction, heightened by recent advances in Large Language Models (LLMs). Its integration with 3D scene understanding is vital for future embodied agents. Past research first explored 3D visual grounding [1, 8, 18, 20] and established new benchmarks including 3D dense captioning [9, 10], open-vocabulary 3D segmentation [15, 33, 48] and detection [31, 61]. This paper focuses on 3D visual grounding first, with plans to expand language annotations and benchmarks in the future. Our visual grounding benchmark aligns with the multi-view setting of the basic 3D perception track, taking multiple ego-centric RGB-D images as input, and includes tenfold more complex prompts.

3. Dataset

This section presents the dataset construction, including data processing and annotation, and shows the statistics.

3.1. Data Collection & Processing

Ego-Centric Sensor Data Collection. Considering there have been readily available 3D indoor scene scans from existing datasets, we start with integrating those providing ego-centric RGB-D captures with the corresponding camera poses. Given the compatibility of ScanNet [13], 3RScan [51], and Matterport3D [6], we select the high-quality part with complete regions and necessary annotations to form the initial version of EmbodiedScan (Fig. 2). ARKitScenes [2], possessing different data organization, sensors, and annotations, is considered for future inclusion.

Frame Selection & Scene Division. Although these datasets all have RGB-D data, the data format, sampling frequency, and relationships among viewpoints are different. See more details in the appendix. We first unified the format into a general multi-view case to fit Matterport3D by adding randomness when loading images but maintaining sequential continuity for ScanNet and 3RScan during inference. Our model can thus handle both temporal and randomly captured multi-view images. Additionally, we divided building-scale scenes of Matterport3D into regions based on official annotation, selecting corresponding images with depth points falling into the region. As for different sampling rates of images in ScanNet and 3RScan videos, we sample one keyframe per 10 frames for ScanNet and keep all the images for 3RScan. The uniform sampling is generally in line with the actual situation.

Global Coordinate System. A global coordinate system is necessary to aggregate multi-view observations and serve as a reference for outputs. We follow the convention of ScanNet, deriving a system with the origin around the center of the scene, the horizontal plane lying on the floor and axes aligning with walls [34]. This post-processing harmonizes the data distribution, slightly improving performance on benchmarks. Practical applications may not have such a prior global system or vary according to observations, posing another interesting problem for future exploration.

3.2. Annotation

We provide three types of annotations - 3D bounding boxes, semantic occupancy, and language descriptions - each serving to enrich different aspects of scene understanding.

3D Bounding Boxes. Following standard definitions [2, 3], a cuboid is defined by its 3D center, size, and ZXY Euler angle orientation. We used the Segment Anything Model (SAM) [22] and a customized annotation tool based on [24] (Fig. 3a) to address limitations in existing 3D box annotations, *i.e.*, lack of orientation and small object annotations. It supports the conventional functionality of annotating 3D boxes with orientation in three orthographic views. Furthermore, we sample several keyframes with clear imaging according to the camera pose changes and ensure they cover non-overlap regions and most objects to generate SAM masks and axis-aligned boxes for further adjustment. We work with an annotation team and check the quality of all the labels in the end. Each scene takes around 10-30 minutes to annotate, varying with the scene complexity.

Semantic Occupancy. Semantic occupancy necessitates accurate boundaries across semantic regions without considering object pose or recalling all the objects, so the original point cloud segmentation annotations were more suitable to be used for deriving occupancy ground truth. For each voxel, we assigned the category with the most points as the semantic label for that cell. A compromise between perception granularity and computational efficiency resulted in $40 \times 40 \times 16$ occupancy maps in the perception range $[-3.2m \sim 3.2m, -3.2m \sim 3.2m, -0.78m \sim 1.78m]$ along the X-Y (horizontal) plane and Z (vertical) axis.

Language Descriptions. Given updated 3D bounding boxes annotated with orientations, we derive the language prompts that describe the spatial relationships among objects following SR3D [1]. They serve as the prompt input to the language-grounded perception models for performing 3D visual grounding. Due to increased object density after annotation, identifying unique objects became more challenging. To overcome this, we combined multiple spatial-relationship prompts to exclusively ground objects. See more samples in the appendix.

3.3. Statistics

Vocabulary Construction. During labeling, we ask annotators to write semantic categories in an open-vocabulary manner. This was efficient and suited the complex, large-vocabulary dataset and can provide natural annotations for future open-world research. To sort out these labels, we used Sentence-BERT [37] to cluster similar categories with text embeddings, match them to WordNet nodes, and finally revise and merge them manually. The vocabulary shares common categories with COCO (50/69 indoor classes) and LVIS (550/1203).

Instance Statistics. We first show the instances of different categories in Fig. 3c. Our dataset contains over 760 categories, covering common objects in our daily life. More than 288 categories have over 10 instances, and around 400 categories have more than 5 instances. These numbers are 20× higher than most previous works with 3D box annotations and 3× higher than ScanNet instance segmentation annotations with more than 5 instances. There is also a notable increase in object numbers of small boxes and different categories (Fig. 3b and 3c). We remove four categories, {wall, ceiling, floor, object} in our 3D detection benchmark and divide the remaining 284 categories into three splits, {head, common, tail} with {90, 94, 100} classes.

Occupancy Statistics. Semantic occupancy statistics (Fig. 3d) reveal the space occupied by different categories, relevant for navigation and motion planning. We chose the first 80 categories for our occupancy prediction benchmark based on distribution and significance in downstream tasks.

Language Prompts Statistics. Generated language prompts following SR3D fall into five types of spatial object-to-object relations: Horizontal Proximity, Vertical Proximity, Support, Allocentric, and Between. If a scene has between 2 and 6 instances of a certain category, these categories are considered valid target categories. If a scene has a single instance of a category, it is selected as the anchor category. The training/validation set contains 801711/168322 language prompts, nearly 10 times larger than the original SR3D datasets (Fig. 3b and Tab. 1).

4. Embodied Perceptron

Given this dataset, we can take multi-modality input, including RGB images, point clouds derived from depth maps as well as language prompts, to extract multi-modal representations and perform different downstream tasks. This section provides a baseline, namely Embodied Perceptron,

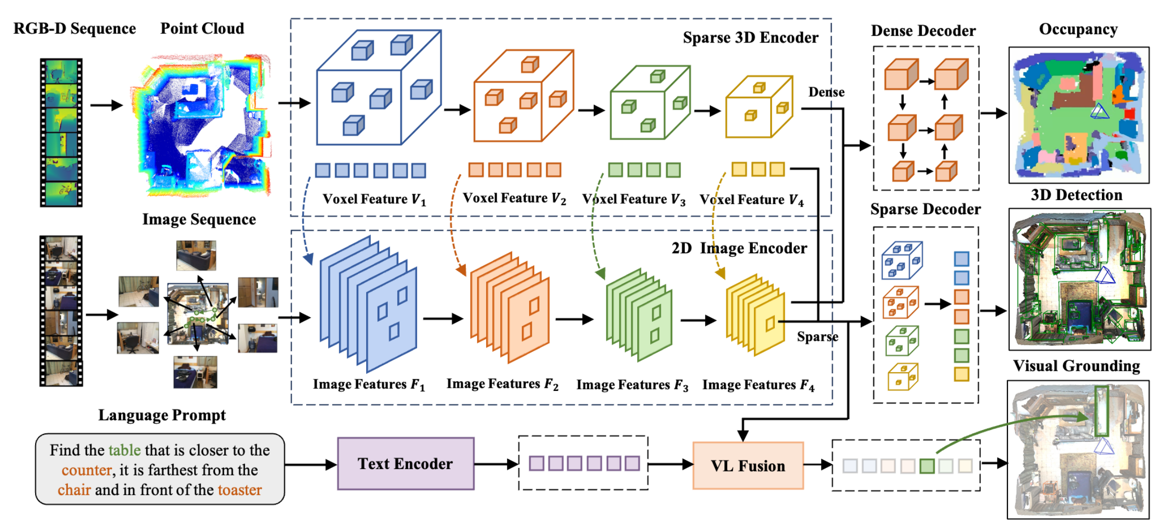


Figure 4. Embodied Perceptron accepts RGB-D sequence with any number of views along with texts as multi-modal input. It uses classical encoders to extract features for each modality and adopts dense and isomorphic sparse fusion with corresponding decoders for different predictions. The 3D features integrated with the text feature can be further used for language-grounded understanding.

with a unified framework and customized design for holistic 3D scene understanding from ego-centric views.

Framework Overview. The framework includes a multi-modal 3D encoder to extract object & scene representations and sparse & dense decoders for various downstream tasks. In addition, we customize the output's parameterization and training objectives to fit the formulation of oriented 3D bounding boxes in the sparse decoder.

4.1. Multi-Modal 3D Encoder

As shown in Fig. 4, the multi-modal 3D encoder first has separate encoders for different modalities - ResNet50 [17] and FPN [27] (optional) for 2D images, Minkowski ResNet34 [11] for point clouds, and BERT [14] for texts. After extracting these features, we further fuse and process them into sparse or dense features for different downstream tasks. Next, we first present how we aggregate multi-view inputs and then introduce different fusion approaches for dense and sparse feature extraction.

Scalability for Input Views. Contrasting with prior works, our framework can accept any number of RGB-D views, making it adaptable and generalizable to varying input orders and quantities. We conveniently aggregate different depth map views by transforming the point clouds into a global coordinate system, downsampling as needed. For multiple images, we query corresponding 2D features using perspective projection from 3D points, averaging them to maintain permutation invariance. This technique allows consistent feature updates during ego-centric exploration. Theoretically, voxel features could be updated by merging the volume feature at frame t with the incremental feature from RGB-D input at frame $t{+}1$. In practice, we accommodate any number of views as batch-wise samples for accelerating training and evaluation. Our model demonstrates notable scalability, where fewer views (e.g., 20) may be used for memory efficiency during training, while more views (*e.g.*, 50) can enhance performance during inference.

Dense Fusion. Previous works typically integrate the color and coordinates of points at the input stage, like "painting" [50], or form multi-modality dense BEV features for concatenated fusion [26, 30]. The latter way suits our occupancy prediction baseline and thus we adopt the straightforward dense fusion on the pre-defined grid, which shares the same resolution with the ground truth. We construct feature volume by projecting grid points onto a *single* 2D feature map post-FPN, then consolidate it with 3D features densified from sparse voxel features. For object detection, we argue that concurrent multi-modality fusion across *multiple* feature levels is more effective.

Isomorphic Multi-Level Multi-Modality Fusion. Formally, the input aggregated points $P \in \mathcal{R}^{N_p \times 3}$ (first voxelized) and N_i images as $I \in \mathcal{R}^{N_i \times H \times W}$ are processed via a Minkowski ResNet and a shared 2D ResNet respectively. This extracts multi-level sparse voxel features $V_k \in \mathcal{R}^{C_k \times N_{V_k}}$ on K levels and image features $F_s \in \mathcal{R}^{C_s \times H_s \times W_s}$ on S levels. In practice, these two ResNets produce 4 levels of features, for both point clouds and images, denoted as isomorphic multi-modality encoders.

In dense fusion, we filter F_s with an upsampling FPN to derive a feature map F_{up} with $stride = 4$ and use it to construct the feature volume for fusing with V_4. For the sparse case, we use multi-level features as seeds instead of a single dense feature map to predict 3D objects. The initial attempt of still query features from F_{up} or raw images I for these seeds is unstable due to inconsistent features for fusion and confusing gradients back-propagation. Thus, we leverage the isomorphic architecture for level-based projec-

Table 2. Continuous and multi-view 3D object detection benchmark on EmbodiedScan (split by the double line).

Methods	Input	Large-Vocabulary				Head		Common		Tail	
		AP_{25}	AR_{25}	AP_{50}	AR_{50}	AP_{25}	AR_{25}	AP_{25}	AR_{25}	AP_{25}	AR_{25}
Camera-Only	RGB	12.80	34.61	4.25	13.07	17.40	44.79	7.64	24.22	0.03	3.09
Depth-Only	Depth	17.16	51.40	10.52	25.75	21.39	61.14	13.27	41.58	2.74	20.91
Multi-Modality	RGB-D	**19.07**	**51.56**	**11.57**	**28.15**	23.54	60.23	15.80	44.99	1.24	17.74
ImVoxelNet [41]	RGB	6.15	20.39	2.41	6.31	10.96	34.29	4.12	15.40	2.63	9.21
VoteNet [34]	Depth	3.20	6.11	0.38	1.22	6.31	12.26	1.81	3.34	1.00	1.83
FCAF3D [40]	Depth	9.07	44.23	4.11	20.22	16.54	61.38	6.73	42.77	2.67	24.83
+our decoder	Depth	14.80	51.18	8.77	27.46	25.98	67.12	10.85	50.08	5.72	32.85
+painting	RGB-D	15.10	**51.32**	8.64	26.66	26.23	67.53	11.39	50.64	5.80	32.13
Ours	RGB-D	**16.85**	51.07	**9.77**	**28.21**	28.65	67.51	12.83	50.46	7.09	31.52

tion and feature fusion, *i.e.*, V_k queries the corresponding image features of F_k, which empirically shows a better and more stable performance. This method enables multi-level multi-modality feature fusion compared to the "painting" approach and ensures the consistency of features and gradients across different network levels and modalities.

Vision-Language (VL) Fusion. Given the multi-level sparse visual features F_k^S and text features from the text encoder, we use a multi-modal fusion transformer model [20, 61] for vision-language information interactions. Each transformer layer uses a self-attention block to refine sparse visual features and exploit spatial relationships. Then visual and text features interact in cross-modal attention blocks. This interaction guides updated sparse grounding features F^G to be context-aware for subsequent prediction.

4.2. Sparse & Dense Decoder

Given multi-modal features from typical encoders, we employ separate fusion streams for sparse and dense tasks. This results in four levels of sparse voxel features F_k^S from isomorphic sparse fusion and a single dense feature F^D for decoding and predictions. These are then processed to obtain 3D box and occupancy predictions.

Sparse Decoder for 3D Boxes Prediction. Using the multi-level fused features F_k^S, we upsample them as in FCAF3D, appending classification, regression, and centerness prediction heads for 3D object detection. In particular, to fit the oriented 3D box output, we add a 6D rotation representation [60] into original regression targets, ultimately decoded as 3D centers $\mathbf{c}$, 3D sizes $\mathbf{l}$, and Euler angles $\mathbf{\Theta}$. Training objectives include the original classification loss, centerness loss, and a disentangled Chamfer Distance (CD) loss for eight corners [3, 44]. Specifically, we use one of three groups of decoded predictions, {3D centers, 3D sizes, and Euler angles}, while setting the other two with ground truths to compute three corner losses. For example, given 3D sizes and Euler angles ground truth, we can derive the corner loss between the predicted $\mathbf{B}$ and the ground truth box $\hat{\mathbf{B}}$ yielded by 3D center prediction errors:

$$L_{\mathbf{c}} = L_{CD}(\mathbf{B}(\mathbf{c}, \hat{\mathbf{l}}, \hat{\mathbf{\Theta}}), \hat{\mathbf{B}}) \tag{1}$$

Together with the corner loss derived by the overall predicted bounding boxes L_{pred}, We balance these losses with preset weights and use them to replace the original box loss:

$$L_{loc} = \lambda_{\mathbf{c}} L_{\mathbf{c}} + \lambda_{\mathbf{l}} L_{\mathbf{l}} + \lambda_{\mathbf{\Theta}} L_{\mathbf{\Theta}} + \lambda_{pred} L_{pred} \tag{2}$$

We set $\lambda_{\mathbf{c}} = \lambda_{\mathbf{l}} = \lambda_{\mathbf{\Theta}} = 0.2$ and $\lambda_{pred} = 0.4$ to highlight the importance of the overall prediction, which performs well empirically. The target assignment strategy and post-processing during inference also follow FCAF3D [40].

Dense Decoder for Occupancy Prediction. With the dense feature F^D, we use a 3D FPN [41] to aggregate multi-level features and produce multi-scale occupancy predictions. Since the task requires more powerful low-level features for fine detail understanding, predictions at each scale are thus supervised with decayed half weights from high to low resolution [5]. We use cross-entropy loss and scene-class affinity loss [55] for training. During inference, we only use the high-resolution output for prediction.

Sparse Decoder for 3D Visual Grounding. Grounding features F^G updated after each transformer layer are fed into the prediction heads sharing the same architecture as those used for 3D detection. All prediction head outputs in each layer are supervised during training for stability and improved performance. An additional contrastive loss aligns the visual feature with target text prompts, ensuring the features of a target text token are closer to corresponding visual features and further from other visual or text tokens.

5. Benchmark

Our benchmark has three categories based on data samples: scene-based, view-based, and prompt-based. Scene-based benchmarks mean the samples are based on different scenes, covering continuous and multi-view perception. View-based benchmarks use ego-centric views for tasks like monocular 3D detection. Lastly, samples of 3D visual grounding are based on constructed language prompts. Detailed splits will be discussed in each benchmark.

For metrics, we use the 3D IoU-based average precision (AP) with thresholds of 0.25 and 0.5 for 3D detection and visual grounding. We also provide average recall (AR) for reference. For occupancy prediction, we employ the mean Intersection of Union (mIoU) as a performance measure. Due to the space limitation, please refer to the appendix for implementation details of different baselines, and more quantitative and qualitative results including an "in-the-wild" evaluation demo.

Table 3. Continuous and multi-view occupancy prediction benchmark (split by the double line). “refri.” means “refrigerator”.

Methods	Input	mIOU	empty	floor	wall	chair	cabinet	door	table	couch	shelf	window	bed	curtain	refri.	plant	stairs	toilet
Camera-Only	RGB	10.43	39.09	34.10	30.24	26.46	9.49	25.53	41.60	35.19	16.22	20.45	24.46	19.80	26.01	17.26	1.83	29.78
Depth-Only	Depth	14.44	73.91	66.22	56.13	49.96	15.70	24.37	56.84	55.35	30.55	26.66	42.81	30.81	33.01	21.71	6.21	45.35
Multi-Modality	RGB-D	**20.79**	73.50	63.64	62.30	54.60	19.96	48.99	61.10	69.76	39.86	34.62	54.83	54.45	48.90	41.22	7.97	63.52
OccNet [43]	RGB	8.07	37.15	46.90	25.63	20.94	13.17	18.40	26.81	22.86	13.59	13.49	26.75	22.92	17.15	17.07	4.77	33.60
SurroundOcc [55]	RGB	9.10	38.54	46.17	23.55	23.04	13.60	19.15	27.79	22.88	13.11	13.72	24.32	18.89	13.58	14.77	7.83	34.71
Camera-Only	RGB	10.48	40.45	41.25	27.19	26.16	15.50	20.30	30.82	26.70	15.01	14.33	29.17	23.30	16.99	15.98	6.17	42.57
Depth-Only	Depth	15.56	69.92	60.52	51.74	49.44	23.08	24.33	45.77	43.52	29.74	23.02	39.04	41.22	17.42	19.58	25.79	60.45
Multi-Modality	RGB-D	**19.97**	71.21	64.92	55.00	52.04	27.35	33.97	47.93	46.26	31.87	27.98	46.58	46.56	24.05	39.01	24.40	67.79

Table 4. Monocular 3D object detection benchmark on EmbodiedScan.

Methods	Input	20 Common Classes				chair	cabinet	table	bin	couch	bed	bathtub	toilet
		AP_{25}	AR_{25}	AP_{50}	AR_{50}								
FCOS3D [52]	RGB	8.93	27.96	0.91	5.00	27.15	1.14	6.21	10.23	9.47	18.38	6.31	40.51
ImVoxelNet [41]	RGB	18.95	52.74	1.81	7.10	46.70	4.63	18.10	17.82	20.39	41.51	10.14	65.70
VoteNet [34]	Depth	14.30	31.44	1.68	5.14	54.00	2.41	19.53	14.72	21.80	45.58	13.49	68.16
ImVoteNet [35]	RGB-D	19.63	34.32	3.88	8.82	56.72	2.88	29.00	21.96	27.77	56.94	37.56	74.08
FCAF3D [40]	Depth	25.70	78.53	5.73	20.26	65.91	6.47	26.64	34.93	22.50	53.68	26.38	71.90
+our decoder	Depth	28.16	84.50	4.92	20.69	63.85	6.62	32.34	38.96	31.61	60.33	38.17	75.57
+painting	RGB-D	30.19	83.93	5.74	21.90	66.39	7.41	33.66	42.86	32.24	60.04	41.31	77.59
Ours	RGB-D	**34.28**	**85.03**	**12.61**	**32.25**	69.47	10.01	37.29	45.17	31.67	63.27	50.63	80.39

5.1. Fundamental 3D Perception Benchmarks

Continuous 3D Perception. As opposed to driving scenarios, indoor scene understanding is typically in an enclosed space, making it important to fully leverage multi-view cues formed by RGB-D sequence and continuously maintain an overall scene-level representation. Thus, we design this new benchmark involving sequential views for perceiving covered 3D regions. Models are trained and evaluated scene-wise with 3930/703/552 scans allocated for training/validation/testing. To accelerate the training and evaluation, we construct N data samples with $1 \sim N$ views from N sampled views per scan. Here, $N = 10$ during training with random view sampling, while in evaluation, $N = 50$ with fixed views. Corresponding instances and occupancy truths are obtained by combining pre-computed visible instance IDs and occupancy masks of selected views. If a category lacks instances, it is removed when calculating mAP and mIoU. Given this new setup, we primarily offer three baselines with different input modalities (Tab. 2 and 3).

Continuous 3D Object Detection. As anticipated, both RGB and depth features significantly impact this task, leading to superior results of our RGB-D approach (Tab. 2). The performance of the depth-only model closely mirrors the multi-modality approach, indicating depth's dominance in 3D perception. Our method of constructing multi-modal features based on sparse voxel features also aligns with this intuition. Low performance on tail categories suggests dataset size influences performance, warranting future enhancement.

Continuous Semantic Occupancy Prediction. This benchmark offers comprehensive results including mIoU and IoU for common classes. Unlike the detection benchmark, there is a notable gap between the depth-only and RGB-D baseline. This might be due to the former's limited semantic understanding capability, especially evident in categories like door and curtain, which are similar to walls in shape. On such a task that requires more fine-grained understanding, the depth sensor's weakness is enlarged. Meanwhile, depth plays a crucial role in predicting empty space, floor, and wall, while RGB information substantially improves prediction for most categories.

Table 5. Multi-view 3D visual grounding benchmark. “Indep/Dep” refer to “View-Independent/Dependent”. Easy/Hard and Indep/Dep have a ratio of 80%/20% and 78%/22%.

Methods	Input	Overall	Easy	Hard	Indep	Dep
		AP_{25}	AP_{25}	AP_{25}	AP_{25}	AP_{25}
ScanRefer [8]	RGB-D	12.85	13.78	9.12	13.44	10.77
BUTD-DETR [20]	RGB-D	22.14	23.12	18.23	22.47	20.98
L3Det [61]	RGB-D	23.07	24.01	18.34	23.59	21.22
Ours	RGB-D	**25.72**	**27.11**	**20.12**	**26.37**	**23.42**

Multi-View 3D Perception. Unlike continuous settings, multi-view 3D perception does not predefine the order of views but provides all views to the model for scene-level results. This setting was studied previously [41], so we first reproduce common methods on our benchmark.

Multi-View 3D Object Detection. We implement baselines including ImVoxelNet [41] with RGB-only input and VoteNet [34] and FCAF3D [40] with depth-only input (Tab. 2). Additional dimensions are added to predict Euler angles with a simple L1 loss on their cosine values, but it yields underwhelming results. Substituting this with our decoder design markedly improves performance. Further using point cloud input painted for FCAF3D with RGB-D input slightly underperforms our baseline. Nevertheless, all models have substantial potential for improvement, demonstrating the challenges of this new dataset and setup.

Multi-View Semantic Occupancy Prediction. We implement two popular baselines from autonomous driving benchmarks, OccNet [43] and SurroundOcc [55] (similar to TPV-Former [19]). Their performance slightly lags behind our camera-only baseline. Variants of our baselines exhibit a performance trend akin to embodied benchmarks.

Monocular 3D Perception. Finally, the basic ego-centric setting is monocular 3D perception, specifically 3D detection, where each data sample comprises a single RGB-D

Table 6. Ablation with conventional settings.

Oriented	Multi-View	AP_{25}	AR_{25}	AP_{50}	AR_{50}
✗	✗	70.17	90.46	54.58	75.66
✓	✗	61.87	90.31	47.30	73.93
✓	✓	59.95	87.92	43.33	69.95

Table 7. Real vs. rendered images on ScanNet.

Train	Val	Overall	Head	Common	Tail
Render	Render	22.11	33.01	16.44	6.74
Render	Real	18.72	27.02	14.85	6.25
Real	Real	21.98	32.91	17.18	5.05

frame and corresponding visible 3D boxes. Scan splits are used to extract frames as data samples, resulting in 689k/115k/86k images for the training/validation/testing.

Monocular 3D Object Detection. This is more challenging than multi-view due to the absence of stereo geometric cues and truncated object views in indoor scenes, so the performance is significantly reduced in large-vocabulary settings. Hence, we first create a benchmark for 20 common categories (Tab. 4), observing a larger AP-AR gap for top methods because of difficulties predicting accurate 3D boxes from partial views. Similarly, our method outperforms others, providing a solid baseline for future studies.

5.2. Language-Grounded Benchmark

Multi-View 3D Visual Grounding. Our benchmark introduces language into the perception loop to foster interactive 3D scene representation learning. With comprehensive instance annotations, our benchmark presents more complex prompts and grounding cases than previous works. As an initial step, this setup takes multi-view RGB-D images as input without considering differing prompt timestamps. The goal is to ground the object described by the language prompt in the scene using information from different ego-centric views. Ground-truth detection boxes are not provided as candidates for grounding during evaluation, which can better validate end-to-end 3D visual grounding ability than the original SR3D [1]. Data sample splits align with previous benchmarks' 3D scan splits.

We reimplement classic methods like ScanRefer [8], BUTD-DETR [20], and L3Det [61] (Tab. 5). Our baseline outperforms all due to the strong multi-modal encoder. However, the performance remains much lower than previous works, partly due to changes in input format and annotations, which we will analyze next. Further challenges arise from handling more categories and small objects, making the grounding task more complex in parsing input prompts and predictions. Addressing these new challenges in this classical task would be promising for future research.

5.3. Analysis

Finally, we make further analysis to connect EmbodiedScan to current progress in computer vision.

Axis-aligned vs. Oriented Boxes. We start with the 18-class detection performance of FCAF3D on ScanNet (Tab. 6). First, we change the annotations to oriented 3D boxes and adapt with our decoder. We find a significant drop in performance, indicating that the orientation estimation makes this task more challenging. We need to explore a better method to represent and predict the object pose.

Table 8. Benefits from training with EmbodiedScan.

Train	Val	Overall	Head	Common	Tail
ScanNet	ScanNet	20.28	29.81	15.57	6.40
Ours	ScanNet	**23.02**	33.82	18.09	6.57
ScanNet	Ours	10.92	21.10	8.06	1.78
Ours	Ours	**16.85**	28.65	12.83	7.09

Reconstructed Point Cloud vs. Multi-View RGB-D. Subsequently, replacing the reconstructed point clouds with the aggregated ones from multi-view depth maps has minor effects on AP_{25} but heavily impacts AP_{50} (Tab. 6), implying that the accuracy of reconstructed point clouds is superior to raw depth maps. Therefore, integrating reconstruction techniques in perception loops shows potential.

Next, we study the gap between the real and rendered images, and the benefits from training with EmbodiedScan. We do the comparison with our multi-modality baseline on the large-vocabulary multi-view 3D detection benchmark.

Real Capture vs. Rendering. As shown in Tab. 7, apart from the significant visual difference between real and rendered images (Fig. 1), the model's performance also has a remarkable decrease when transferring models trained with rendered images to the real world, particularly when the annotations are sufficient (5.99% AP drop on head categories and 5.89% AP lower than models trained with real images.)

Benefits from EmbodiedScan. Finally, we also quantitatively evaluate the benefits of training models with our large-scale EmbodiedScan (Tab. 8). As expected, when training our models with EmbodiedScan, we observed a significant improvement in both ScanNet (2.74% AP) and the overall validation split (5.93% AP), particularly 4.01% AP and 7.55% AP increase on head categories.

6. Conclusion

This paper introduces EmbodiedScan, a multi-modal perception suite aiming for language-grounded holistic 3D scene understanding from ego-centric views. We construct a large-scale dataset with diverse sensor data and multi-modal annotations, including 3D oriented boxes, semantic occupancy and language descriptions. Based on this dataset, we propose a baseline framework capable of handling any number of views input, using a unified multi-modal encoder and task-specific decoders. We establish benchmarks for basic and language-grounded 3D perception. Experiment results highlight our work's value and reveal new challenges in this setup. We believe EmbodiedScan can bring opportunities in embodied 3D perception and may also have a broader impact in related fields with the massive data and rich annotations.

Acknowledgements. This work is supported by Shanghai Artificial Intelligence Laboratory.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, 2020. 3, 4, 8

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2, 3, 4

[3] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4, 6

[4] 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. *CoRR*, abs/1903.11027, 2019. 2

[5] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3, 6

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1, 2, 3

[7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 2

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 2020. 2, 3, 7, 8

[9] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 3

[10] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[11] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5

[12] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 1, 2

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[15] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[18] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[19] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3, 7

[20] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, 2022. 3, 6, 7, 8

[21] Mukul Khanna*, Yongsen Mao*, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. *arXiv preprint*, 2023. 2

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 4

[23] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[24] E Li, Shuaijun Wang, Chengyang Li, Dachuan Li, Xiangbin Wu, and Qi Hao. Sustech points: A portable 3d point cloud interactive annotation platform system. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020. 4

[25] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[26] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 2022. 3, 5

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[28] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020. 3

[29] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[30] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 3, 5

[31] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. *arXiv preprint arXiv:2207.01987*, 2022. 3

[32] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 2

[33] 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*, 2023. 3

[34] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 3, 4, 6, 7

[35] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 3, 7

[36] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 2

[37] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 4

[38] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2

[39] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[40] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, 2022. 1, 3, 6, 7

[41] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*, pages 2397–2406, 2022. 3, 6, 7

[42] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[43] Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 7

[44] Andrea Simonelli, Samuel Rota Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *IEEE International Conference on Computer Vision*, 2019. 3, 6

[45] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2

[46] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[47] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[48] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 3

[49] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 3

[50] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 3, 5

[51] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance relocalization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 2, 3

[52] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021. 3, 7

[53] Tai Wang, Jiangmiao Pang, and Dahua Lin. Monocular 3d object detection with depth from motion. In *European Conference on Computer Vision (ECCV)*, 2022.

[54] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 3

[55] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 6, 7

[56] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

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

[58] Wenwei Zhang, Zhe Wang, and Chen Change Loy. Exploring data augmentation for multi-modality 3d object detection. *arXiv preprint arXiv:2012.12741*, 2020. 3

[59] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[60] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 6

[61] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*, 2023. 3, 6, 7, 8