

# Humanoid Occupancy: Enabling A Generalized Multimodal Occupancy Perception System on Humanoid Robots

Wei Cui<sup>\*1</sup>, Haoyu Wang<sup>\*2</sup>, Wenkang Qin<sup>2</sup>, Yijie Guo<sup>1</sup>, Gang Han<sup>1</sup>, Wen Zhao<sup>1</sup>, Jiahang Cao<sup>1</sup>, Zhang Zhang<sup>1</sup>, Jiaru Zhong<sup>1</sup>, Jingkai Sun<sup>1</sup>, Pihai Sun<sup>1</sup>, Shuai Shi<sup>1</sup>, Botuo Jiang<sup>1</sup>, Jiahao Ma<sup>1</sup>, Jiaxu Wang<sup>1</sup>, Hao Cheng<sup>1</sup>, Zhichao Liu<sup>2</sup>, Yang Wang<sup>2</sup>, Zheng Zhu<sup>2</sup>, Guan Huang<sup>2</sup>, Jian Tang<sup>†1</sup> and Qiang Zhang<sup>†Δ1</sup>

<sup>\*</sup> Equal Contributors, <sup>†</sup> Corresponding Authors, <sup>Δ</sup> Project and Technical Leader

<sup>1</sup> X-Humanoid, <sup>2</sup> GigaAI

Project Page: <https://humanoid-occupancy.github.io>



Figure 1: Schematic diagram of the Humanoid Occupancy system.

## Abstract

Humanoid robot technology is advancing rapidly, with manufacturers introducing diverse heterogeneous visual perception modules tailored to specific scenarios. Among various perception paradigms, occupancy-based representation has become widely recognized as particularly suitable for humanoid robots, as it provides both rich semantic and 3D geometric information essential for comprehensive environmental understanding.

In this work, we present Humanoid Occupancy, a generalized multimodal occupancy perception system that integrates hardware and software components, data acquisition devices, and a dedicated annotation pipeline. Our framework employs advanced multi-modal fusion techniques to generate grid-based occupancy outputs encoding both occupancy status and semantic labels, thereby enabling holistic environmental understanding for downstream tasks such as task planning and navigation. To address the unique challenges of humanoid robots, we overcome issues such as kinematic interference and occlusion, and establish an effective sensor layout strategy. Furthermore, we have developed the first panoramic occupancy dataset specifically for humanoid robots, offering a valuable benchmark and resource for future research and development in this domain. The network architecture incorporates multi-modal feature fusion and temporal information integration to ensure robust perception. Overall, Humanoid Occupancy delivers effective environmental perception for humanoid robots and establishes a technical foundation for standardizing universal visual modules, paving the way for the widespread deployment of humanoid robots in complex real-world scenarios.

## 1. Introduction

Humanoid robots are considered the most promising general-purpose robots for integration into human environments due to their human-like structure and movement (Hirai et al., 1998). Despite rapid progress in AI and robotics, their visual perception systems still face key challenges (Bonci et al., 2021). These include designing optimal multimodal sensor layouts for complex environments, efficiently collecting and annotating large-scale diverse datasets, and achieving generality and standardization in visual system design.

From a task-driven perspective, humanoid robots face three core tasks: **manipulation, locomotion, and navigation** (Gu et al., 2025). Manipulation tasks require rich texture and geometric information for precise object recognition and handling; locomotion tasks emphasize terrain geometry perception to ensure stable walking in complex environments; navigation tasks demand a global understanding of environmental semantics and spatial geometry for efficient path planning and autonomous navigation. These diverse requirements present significant challenges for the design of perception systems (DiSalvo et al., 2002; Roychoudhury et al., 2023).

In this work, we focus on enhancing the environmental perception and navigation capabilities of humanoid robots. Environmental perception forms the basis for executing complex tasks, while navigation remains a core challenge in humanoid robotics, with notable advances in areas such as autonomous driving and aerial robotics. However, the unique structural characteristics of humanoid robots introduce challenges such as kinematic interference, data scarcity, and limited representational capacity, which continue to restrict their effective deployment in real-world environments.

We propose the Humanoid Occupancy system (Fig. 1) to provide humanoid robots with a unified, efficient, and information-rich environmental perception capability. The choice of occupancy as the core visual representation paradigm is motivated by several considerations. First, occupancy can directly encode the occupancy status of each spatial unit in the environment in a voxel or grid format, capturing not only the distribution on the 2D plane but also detailed structural and semantic attributes along the vertical dimension. This makes it far superior to traditional representations such as Bird’s Eye View (BEV) (Phillion and Fidler, 2020), which typically focus only on ground projections and fail to reflect vertical structures and high-level semantic information.



Second, occupancy representation (Elfes, 2013) is naturally suited for multimodal data fusion, allowing information from RGB, depth, LiDAR, and other sensors to be unified within a spatial grid for comprehensive environmental understanding. Compared to other 3D representations such as point clouds (Rusu and Cousins, 2011) or meshes, occupancy offers greater generality and extensibility in data structure, semantic annotation, and downstream task interfaces, facilitating large-scale data processing and model training. Moreover, the dense spatial distribution output by occupancy directly supports path planning, obstacle avoidance, and manipulation tasks, significantly enhancing the adaptability of humanoid robots in complex environments. Therefore, occupancy representation provides a solid foundation for efficient perception across multiple tasks and scenarios.

In terms of system implementation, we adopt a standard three-stage visual system architecture, comprising hardware design, dataset construction, and a multimodal fusion network. We propose an innovative sensor layout strategy tailored for humanoid robots, effectively mitigating perception blind spots caused by structural interference. Additionally, we have constructed the first panoramic occupancy dataset specifically for humanoid robots, providing a valuable benchmark and resource for future research and applications. Finally, we have integrated and tested the system on the Tienkung humanoid robot platform, and experimental results demonstrate its superior environmental perception and navigation performance in complex environments.

## 2. Preliminaries and Related Works

### 2.1. Perception System of Humanoid Robots

The development of visual perception systems for humanoid robots has evolved from single-sensor configurations to multimodal fusion architectures (Ackerman). Early systems primarily relied on monocular or stereo cameras (Ben et al., 2025; Cheng et al., 2024; Fu et al., 2024; Hirose and Ogawa, 2007; Lu et al., 2024; Mutlu et al., 2006; Peng et al., 2025; Shigemi, 2017), but their limited perception range and depth sensing capabilities made it difficult to meet the demands of complex scenarios. With advances in sensor technology, RGB-D cameras (Duan et al., 2024; Long et al.; Sun et al., 2025; Ze et al., 2024; Zhang et al., 2025; Zhuang et al., 2024), LiDAR (Wang et al., 2025; Zhang et al., 2025), panoramic cameras (Zhang et al., 2025), and other sensor types (He et al., 2025; Kim et al., 2024) have been increasingly integrated into robotic platforms, laying the foundation for more comprehensive environmental perception in humanoid robots.

Current mainstream visual perception systems for humanoid robots can be categorized into the following technical paradigms:

- **Vision-Based Perception Systems**

These systems are centered around camera arrays and leverage computer vision and deep learning algorithms for object detection, semantic segmentation, and 3D reconstruction. A typical example is Boston Dynamics' Atlas robot (Ackerman), which is equipped with multiple camera modules for environmental understanding and navigation. In recent years, breakthroughs in Transformer architectures and self-supervised learning have significantly improved the generalization capabilities of these systems in object recognition and scene understanding tasks (Carion et al., 2020; Kirillov et al., 2023; Radford et al., 2021).

- **Multimodal Fusion Perception Systems**

By fusing RGB images, depth data, LiDAR point clouds, and other sources, these systems exploit the complementarity of different sensors to enhance perception accuracy and robustness. For example, Agility Robotics' Digit robot (AgilityRobotics) employs a fusion scheme of LiDAR and vision, enabling stable locomotion and obstacle detection in complex terrains. Multimodal fusion not only improves the perception of dynamic obstacles and textureless regions but also enhances adaptability under varying lighting conditions. Current research focuses on deep learning-based multimodal feature fusion networks (Rudin et al., 2025), which effectively integrate heterogeneous sensor information to further optimize the

accuracy of environmental modeling.

- **End-to-End Perception and Decision-Making Systems**

These frameworks utilize deep neural networks to integrate perception, understanding, and decision-making, directly outputting motion control commands (Luo et al., 2024; Wang et al., 2025). While they demonstrate strong adaptability in specific tasks, they require large-scale datasets and significant computational resources. Recently, reinforcement learning and imitation learning have been widely applied in end-to-end systems, advancing the intelligence of humanoid robots in autonomous navigation and manipulation tasks.

- **Spatial Representation-Based Perception Systems**

Spatial representation methods such as occupancy grids, point clouds, and meshes have become core technologies for providing structured environmental information to downstream tasks (e.g., navigation, manipulation) (Zhang et al., 2025). Among them, occupancy representations (Sima et al., 2023), which encode both rich semantic and geometric information, have emerged as a key direction for enhancing environmental understanding. These methods support 3D map construction and dynamic scene modeling, and provide efficient interfaces for path planning and obstacle avoidance. Mainstream techniques include voxel-based dense reconstruction and neural implicit field-based scene modeling, which have greatly advanced high-precision perception of complex environments.

Notably, the emergence of novel spatial representation methods such as 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023; Wang et al., 2024,) and Neural Radiance Fields (NeRF) (Mildenhall et al., 2021; Ze et al., 2023) has further improved the expressive power and rendering efficiency of spatial representations. 3DGS achieves fast rendering and reconstruction of large-scale scenes through efficient Gaussian distribution modeling, while NeRF employs neural networks for implicit modeling, enabling high-quality 3D reconstruction and novel view synthesis. These methods offer new technical pathways for real-time perception and interaction in complex environments, expanding the application boundaries of spatial representations in robotics.

With advances in computing hardware and algorithms, real-time perception and large-scale data processing capabilities have improved significantly. The emergence of open-source platforms such as ROS (Macenski et al., 2023, 2022; Stanford Artificial Intelligence Laboratory et al.) and Open3D (Zhou et al., 2018), as well as standardized datasets, has accelerated the iteration and cross-platform application of visual perception systems. Current research also focuses on key technologies such as sensor self-calibration, self-supervised data annotation, and cross-domain generalization to further enhance the practicality and robustness of these systems. Overall, visual perception systems for humanoid robots are evolving toward multimodal fusion, optimized spatial representation, and integrated intelligent decision-making. How to efficiently arrange sensors within limited space and achieve effective data fusion and representation has become a core issue for improving system performance, laying a solid foundation for the introduction of occupancy representations and subsequent system design.

Building on these paradigms, this work proposes a novel occupancy perception framework for humanoid robots, which integrates multimodal sensors through a front-end fusion architecture. This framework adopts a annotation pipeline and incorporates temporal feature fusion to enhance dynamic scene understanding, aiming to address key challenges in practical applications.

## 2.2. Occupancy Perception

Occupancy perception has emerged as a key paradigm in 3D scene understanding, especially in autonomous driving, where occupancy networks and grid-based representations enable robust semantic mapping and navigation. These advances have demonstrated the power of dense spatial reasoning, multimodal fusion, and real-time inference for complex environments. However, the application of occupancy-based methods in humanoid robotics remains limited, despite their unique suitability for human-centric, dynamic scenarios.

For humanoid robots, occupancy perception offers distinct advantages: it supports fine-grained spatial reasoning



in near-field environments, which is essential for precise foot placement, manipulation guidance, and safe navigation in cluttered spaces. Unlike vehicles, humanoid robots require high-resolution occupancy maps within short ranges and must address frequent occlusions and articulated motions. This creates new challenges for sensor layout, data annotation, and multimodal fusion, demanding solutions tailored to the robot’s kinematics and operational context.

Recent research has begun to adapt occupancy networks for robotic platforms, integrating RGB, depth, and LiDAR sensors to construct unified occupancy grids, developing annotation pipelines for dynamic scenes, and designing fusion networks that leverage temporal and semantic cues (Wang et al., 2025; Wu et al., 2024). These efforts aim to provide holistic environmental understanding for manipulation, obstacle avoidance, and autonomous navigation (Zhang et al., 2025). Nevertheless, robust multimodal fusion, real-time inference, and scalable dataset construction remain open challenges, compounded by the lack of standardized benchmarks for humanoid robots.

We can broadly categorize prior work on occupancy according to our requirements into three main types:

- **Occupancy Perception Datasets**

Recent advances in 3D occupancy prediction heavily rely on large-scale annotated datasets. OpenOccupancy (Wang et al., 2023) pioneers the first benchmark for surround-view occupancy perception, providing dense voxel-level annotations for LiDAR and camera data. Similarly, Occ3D (Tian et al., 2023) introduces a label generation pipeline that produces dense, visibility-aware labels for any given scene. However, these datasets focus on autonomous driving scenarios with long-range perception (50-200m) and meter-level resolution, neglecting two critical aspects of humanoid robotics: (1) Near-field high-resolution demands – requiring decimeter-level accuracy in 10m range for foot placement; (2) Unstructured human interactions – needing dense annotations for articulated motions and dynamic scenarios.

- **Vision-Based Occupancy Prediction**

Pure camera-based approaches have evolved rapidly by leveraging bird’s-eye-view (BEV) representations (Ma et al., 2024). TPVFormer (Huang et al., 2023) proposes a tri-perspective view decomposition for efficient 3D scene modeling from 2D images. FB-Occ (Li et al., 2023) introduces a forward-backward projection mechanism to enhance occlusion reasoning, while FlashOcc (Yu et al., 2023) achieves real-time performance through compressed volumetric features. Complementary works like VoxFormer (Li et al., 2023) and OccFormer (Zhang et al., 2023) demonstrate transformer-based voxel prediction advancements. SparseOcc (Tang et al., 2024) utilizes mask-guide sparse sampling to enable sparse queries to optimize efficiency. Despite their progress, these methods often struggle with depth ambiguity in cluttered environments.

- **Multimodal Fusion for Occupancy**

Multisensor fusion methods combine complementary sensor strengths. BEVFusion (Liang et al., 2022; Liu et al., 2022) unifies LiDAR and camera features through late fusion in BEV space, achieving superior performance on nuScenes (Caesar et al., 2020). OccFusion (Ming et al., 2024; Zhang et al., 2024) extends this with voxel-based cross-modal attention, enabling dense occupancy prediction from asynchronous sensor streams. These frameworks establish strong baselines for multimodal 3D perception.

### 3. Humanoid Occupancy

#### 3.1. System Overview

The design of perception systems for humanoid robots must simultaneously address the requirements of manipulation, locomotion, and navigation. These three core tasks often impose conflicting constraints, necessitating careful trade-offs in sensor selection and placement. For example, mounting the sensor suite on the chest is advantageous for navigation and locomotion, providing a stable and unobstructed field of view for mapping and obstacle avoidance. However, this configuration can lead to occlusions and interference during tabletop

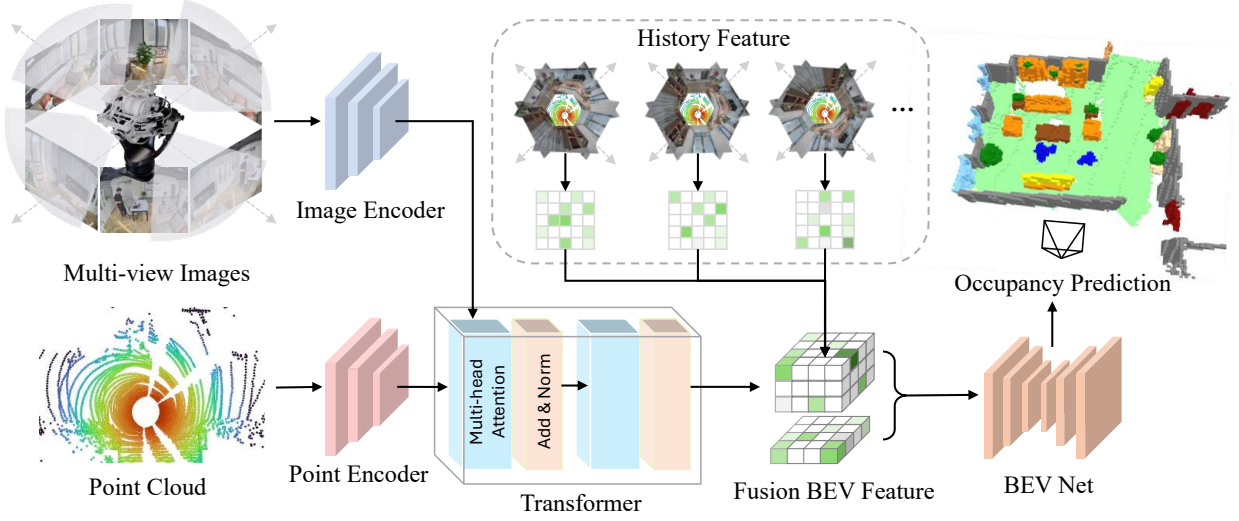


Figure 3: **Overview of our proposed HumaniodOcc.** The proposed network processes image and LiDAR inputs through separate encoders to extract features. The image features and LiDAR features are then fused via cross-attention mechanisms, enabling adaptive interaction between visual and geometric cues. The fused features undergo temporal fusion to aggregate sequential information. Finally, the network predicts the 3D occupancy grid using a BEV encoder and decoder head. More details in [Sec. 3.4](#).

manipulation or when handling objects such as boxes, limiting the robot’s operational versatility.

The development of visual sensor technology for humanoid robots remains an ongoing process, with many sensors imposing strict structural requirements. For instance, LiDAR sensors are typically unsuitable for installation on moving, non-rigid joints unless highly accurate IMU devices are available to provide precise pose and location. This constraint further complicates sensor layout, as designers must balance the need for comprehensive environmental coverage with the mechanical limitations of the robot’s kinematic structure.

In this work, we implement the Humanoid Occupancy concept system on the Tienkung humanoid robot. The system features a modular, selectable RGB-D camera with two degrees of freedom (pitch and yaw), supporting both manipulation and terrain perception tasks. Additionally, 6 cameras and one LiDAR sensor are deployed to enable robust mapping and semantic understanding of the surrounding environment. This configuration is designed to maximize the robot’s perceptual capabilities across diverse operational scenarios, while minimizing interference and ensuring adaptability for future sensor upgrades.

### 3.2. Sensor Layout and Data Acquisition

**Sensor Layout** Our sensor consists of 6 cameras and a LiDAR.

The 6 cameras use standard RGB sensors, arranged in a way that one is arranged in the front and back, and two are arranged on each side. The horizontal FOV of the camera is 118 degrees and the vertical FOV is 92 degrees. The LiDAR uses a 40-line 360-degree omnidirectional LiDAR with a vertical FOV of 59 degrees. The final hardware device structure can be referred to [Fig. 4](#).

**Data Acquisition** When collecting data for autonomous driving perception models, vehicles are usually used to collect data directly on the road. But when it comes to humanoid occupancy data acquisition, it is hard to



Figure 2: **Equipment and Schematic Diagram of Data Acquisition Process.**





Figure 4: **Structural visualization of humanoid occupancy hardware.** We equipped the humanoid robot head with 6 cameras and a lidar to acquire sensory information. More details in [Sec. 3.2](#)

use humanoid robotics to collect data in the environments, due to the extremely high cost and considerable difficulty. we use a wearable device to overcome the above challenges. Our device is constructed using the same sensor configuration as that used on humanoid robots as shown in [Fig. 2](#). So human data collectors can wear the device directly on their heads to collect data in the environment. This collection method greatly reduces our collection cost and difficulty.

In order to ensure that the distribution of collected data is as close as possible to the distribution of real robot data, we require the height of human data collectors to be around 160 cm. This ensures that the height of the sensors worn by human collectors is almost the same as the height of the sensors finally installed on the humanoid robot. To prevent the head shaking of human collectors during collection, causing the sensor to lose horizontal stability, we also added a neck stabilizer to the collection equipment.

### 3.3. Annotation Pipeline

We divide the collected data into three categories according to the scenes, namely home scenes, industrial scenes and outdoor scenes. In the three different scenes, we define different point-wise semantic categories to be annotated. We annotate the bounding boxes of dynamic objects, including pedestrians, cyclists and vehicles. Pedestrians are non-rigid targets, and sometimes other object points are included when they are annotated using bounding boxes. To solve this problem, we divide pedestrians into special posture pedestrians, that is, pedestrians that cannot be represented by bounding boxes, and ordinary posture pedestrians, that is, pedestrians that can be represented by bounding boxes. Ordinary posture pedestrians only require bounding box annotation, while special posture pedestrians require not only bounding box annotation but also point-by-point annotation within the bounding box to distinguish pedestrians from other objects.

In addition to the bounding boxes of dynamic targets, we also perform point-by-point semantic annotation on the point cloud. Our definition of semantic categories to be annotated in different scenarios can be found in the [Tab. 1](#). When performing semantic annotation of point clouds, we remove the dynamic targets of each clips and then superimpose the remaining static points on multiple frames of point clouds. Then, point-level semantic annotation is performed by referring to the projection of the point cloud on the image.

The process of obtaining the occupancy ground truth after annotation can be referred to [Fig. 5](#). After obtaining the static superimposed point cloud annotation and dynamic target box, we first align the superimposed static background points to the ego coordinate system of the frame-by-frame point cloud, and then splice the superimposed dynamic foreground points into the point cloud according to the dynamic target posture of the frame. The superimposed and stitched point cloud is directly voxelized to obtain the final occupancy ground

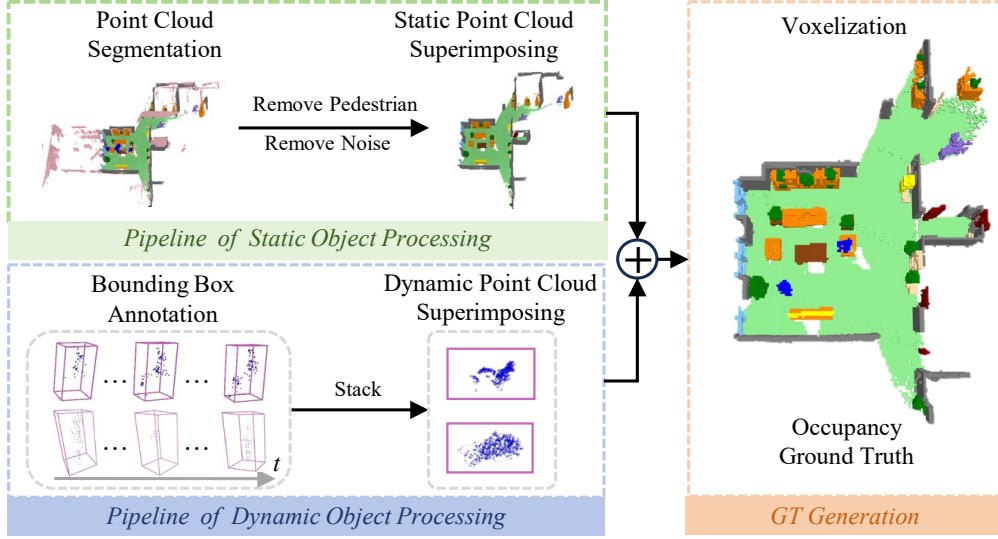


Figure 5: **The occupancy generation pipeline.** The ground truth generation pipeline begins with 3D bounding box annotation and semantic segmentation labeling. Multi-frame LiDAR scans are then processed by separately stitching dynamic objects through bounding box alignment and accumulating static scenes via motion alignment. Finally, the merged points are voxelized to produce the 3D occupancy grid with both occupancy status and semantic labels. More details in [Sec. 3.3](#).

truth. We do not perform Poisson reconstruction on the superimposed and stitched point cloud.

### 3.4. Multi-Modal Fusion Network

Our occupancy perception model accepts multimodal inputs, including a LiDAR point cloud and 6 pinhole camera images. We use the Bird’s Eye View (BEV) paradigm that has been widely validated and adopted in autonomous driving for feature extraction and feature fusion. Since the robotic sensors undergo pitch and roll motions during movement, it is essential to transform the sensor data into a gravity-aligned egocentric reference frame to comply with the BEV assumption. Specifically, we extract LiDAR and camera features through two modality-specific feature extraction branches, and then perform multi-modal feature fusion through Transformer Decoder. The final occupancy result is predicted on the fused BEV features. The overall model architecture can be referred to [Fig. 3](#).

**Camera Feature Encoder** In the image feature extraction branch, we use convolutional neural networks to extract features from camera images. The images of the 6 cameras are extracted from each image through a shared backbone. It is worth noting that our input images are distorted pinhole camera images, and we do not perform dedistortion processing to reduce the overall latency of the system. We made adjustments in the subsequent multimodal feature fusion part so that the distorted image results will not affect the final occupancy prediction.

**LiDAR Feature Encoder** In the LiDAR feature extraction branch, we use PointPillar ([Lang et al., 2019](#)) to construct BEV features. The LiDAR point cloud is first voxelized into a set of non-empty pillars. Each pillar aggregates raw points and encodes their attributes such as coordinates and reflectance via a simplified PointNet-style MLP to extract pillar-wise features. These features are then scattered back to their original pillar locations to form a sparse pseudo-BEV feature map. A 2D convolutional backbone further processes the sparse features into a dense BEV representation, capturing multi-scale spatial context for downstream fusion.

**Multimodal Fusion** When performing multimodal feature fusion, we mainly refer to the DeepFusion ([Li et al., 2022](#)) approach. We perform multimodal fusion through LiDAR to camera cross attention. The LiDAR BEV features extracted by PointPillar are used as queries, and the 6 cameras convolution features are used as keys



Table 1: **Point semantic categories.** We used different sets of semantic labels in three different scenarios.

Home scenes							
1	2	3	4	5	6	7	8
pedestrian	robot	chair	table	floor	wall	window	door
9	10	11	12	13			
plant	appliance	furniture	objects	other			
Industrial scenes							
1	2	3	4	5	6	7	8
pedestrian	floor	wall	conveyor	static objects	dynamic objects	robot	others
Outdoor scenes							
1	2	3	4	5	6	7	8
pedestrian	bicycle	vehicle	road	curb	building	pole	tree
9	10	11	12				
shrub	grass	obstacle	others				

and values. In order to more accurately describe the projection relationship from LiDAR to camera, we use Deformable Attention (Zhu et al., 2021) for feature sampling. When sampling camera features from LiDAR queries reference points projection, we use pinhole camera distortion projection to reduce or eliminate the impact of distorted image feature extraction. LiDAR features provide strong geometric information, and camera features supplement semantic information. The fused features have rich geometric and semantic features.

**Historical Feature Fusion** To leverage temporal information while maintaining computational efficiency, we adopt a concise yet effective approach for fusing historical BEV features. Following the paradigm established in BEVDet4D (Huang and Huang, 2022), we maintain a dynamic feature queue storing BEV features from previous timesteps. For each historical BEV feature map,

$$F_{t-\Delta t} \in \mathbb{R}^{H \times W \times C} \quad (1)$$

where  $\Delta t$  represents the time difference, we first align it to the current ego coordinate frame using the relative ego-motion transformation  $T_{t \rightarrow (t-\Delta t)}$ :

$$F_{t-\Delta t}^{aligned} = \mathcal{W}(F_{t-\Delta t}, T_{t \rightarrow (t-\Delta t)}) \quad (2)$$

where  $\mathcal{W}(\cdot)$  denotes the bilinear interpolation-based warping operation that transforms features according to the ego-motion. The aligned historical features are then concatenated with the current frame’s BEV features along the channel dimension:

$$F_t^{temp} = \text{Concat}([F_t, F_{t-1}^{aligned}, \dots, F_{t-k}^{aligned}]) \quad (3)$$

To effectively fuse these multi-temporal features, we employ BEV encoder to further encodes the feature in the BEV space. We utilize ResNet with classical residual block to construct the backbone and combine the features with different resolutions by applying FPN (Lin et al., 2017).

**Prediction Head** Following FlashOcc’s (Yu et al., 2023) efficient design, we transform 2D BEV features into 3D voxel space through channel-to-height redistribution. FPN features are upsampled and merged through concatenation and convolution. The consolidated BEV feature map is reshaped into 3D voxels by splitting channels into predefined height bins. Finally, a 3D convolution block processes the 3D voxel features with minimal computational cost. To train the proposed baselines, focal loss  $\mathcal{L}_{focal}$  and lovasz-softmax loss  $\mathcal{L}_{ls}$  are leveraged to optimize the network. Following Monoscene, we also utilize affinity loss  $\mathcal{L}_{scal}^{geo}$  and  $\mathcal{L}_{scal}^{sem}$  to optimize the scene-wise and class-wise metrics (i.e., geometric IoU and semantic mIoU). Therefore, the overall

Table 2: **3D semantic occupancy prediction performance on our dataset.** For a fair comparison, we train all models on our dataset with same setups.

Method	Settings		Params	Metrics		Class Metrics											
	Modality	Frames		mIoU	rayIoU	pedestrian	robot	chair	table	floor	wall	window	door	plant	appliance	furniture	objects
BEVDet (Huang et al., 2021)	C	1	75.3M	47.93	59.13	39.85	55.93	43.31	47.76	60.79	51.91	37.77	44.27	60.25	50.89	54.76	27.66
FBOcc (Li et al., 2023)	C	1	76.8M	47.37	59.10	39.78	55.43	42.95	45.37	62.69	51.19	35.73	44.57	59.12	49.36	53.99	28.27
BEVFusion (Liu et al., 2022)	C+L	1	60.6M	53.98	60.24	51.00	51.70	36.72	43.31	66.64	63.38	41.42	71.40	65.29	61.38	62.96	32.45
<b>HumaniodOcc (Ours)</b>	C+L	1	<b>40.5M</b>	<b>52.79</b>	<b>60.49</b>	49.23	49.22	35.28	43.63	57.66	63.90	43.07	71.41	64.17	60.89	59.43	35.54
BEVDet (Huang et al., 2021)	C	2	75.3M	47.95	55.00	39.37	55.66	28.32	39.29	62.42	51.56	36.85	58.61	63.85	54.96	55.03	29.52
FBOcc (Li et al., 2023)	C	2	76.8M	46.70	58.45	40.11	53.52	42.97	45.44	62.14	50.83	35.38	41.06	58.08	49.07	54.00	27.76
BEVFusion (Liu et al., 2022)	C+L	2	60.6M	53.21	59.56	51.00	52.16	37.03	40.83	57.63	63.88	44.38	71.68	63.87	62.45	61.16	32.43
<b>HumaniodOcc (Ours)</b>	C+L	2	<b>40.5M</b>	<b>55.73</b>	<b>61.32</b>	50.85	56.16	40.49	45.49	65.60	64.76	44.07	72.77	65.58	63.01	64.50	35.48

loss function can be derived as:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \mathcal{L}_{ls} + \mathcal{L}_{scal}^{geo} + \mathcal{L}_{scal}^{sem} \quad (4)$$

## 4. Experiments and Results

### 4.1. Implementation details

**Dataset** Our experiments utilize a multi-modal dataset collected from 6 cameras and 1 LiDAR sensor, comprising 180 clips as training set and 20 clips as evaluation set. Each clip contains 200 frames, ensuring rich temporal context for dynamic scene understanding. The LiDAR data provides 3D bounding box annotations and point-wise semantic segmentation labels, which are leveraged to generate high-quality voxelized occupancy ground truth.

**Metrics** We evaluate our occupancy predictions using two metrics. We use mIoU to measure semantic segmentation accuracy in 3D space, computed as the mean IoU across all classes. Following SparseOcc (Liu et al., 2023), we use rayIoU by computing IoU along sampled lidar rays to solve the inconsistency penalty along the depth axis.

**Network settings** The occupancy prediction range is set to [-10m, 10m] for the X and Y axes, and [-1.5m, 0.9m] for the Z-axis, all within the ego coordinate system. The final output 3D occupancy grid has a shape of  $200 \times 200 \times 24$ , with each voxel measuring [0.1m, 0.1m, 0.1m]. For the camera branch, we adopt a pretrained ResNet50 with FPN as the 2D backbone, where input multi-view images are resized to  $960 \times 768$ .

**Training details** Our model is implemented based on mmdetection3d framework and trained on 8 NVIDIA A100 GPUs with a batch size of 4. We adopt the AdamW optimizer with a cosine annealing learning rate scheduler, incorporating a warmup phase, and set the initial learning rate to  $2e-4$ . The model is trained for 20 epochs using a StreamPETR-style (Wang et al., 2023) continuous frame training strategy and temporal information is incorporated starting from the second epoch.

### 4.2. Main results

We benchmark our method against representative BEV perception models on our multi-modal dataset. Our comparative analysis evaluates occupancy prediction performance both on single-frame settings and multi-frame settings with historical information. The comparison between our model and some general methods can be seen in Fig. 7. All models adopt identical training configurations, including input image resolution, backbone network, feature dimensions, and training strategies. As shown in Tab. 2, our model achieves superior metrics while maintaining lightweight architecture with significantly fewer parameters. We also show the visualization results of our perception system in Fig. 6.

### 4.3. Ablation study

Ablation studies are conducted to validate the effectiveness of components in our approach. We analyze the impacts of camera distortion correction strategies, temporal multi-frame aggregation, and different sensor



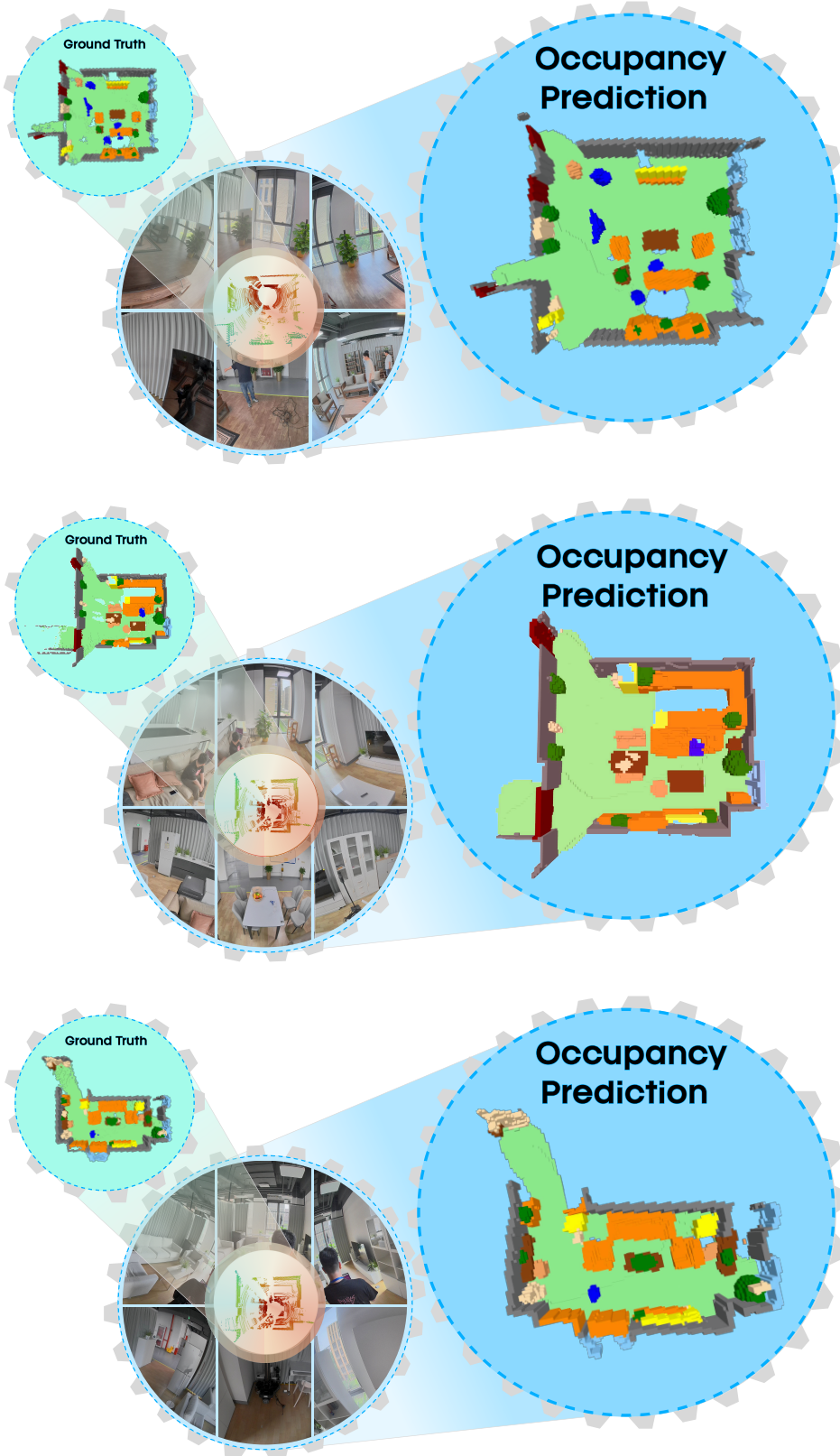


Figure 6: **The results of our humanoid perception system.** The left side of each row shows the occupancy ground truth, the middle is the 6 images and point cloud input to the model, and the right side shows the result of model inference.

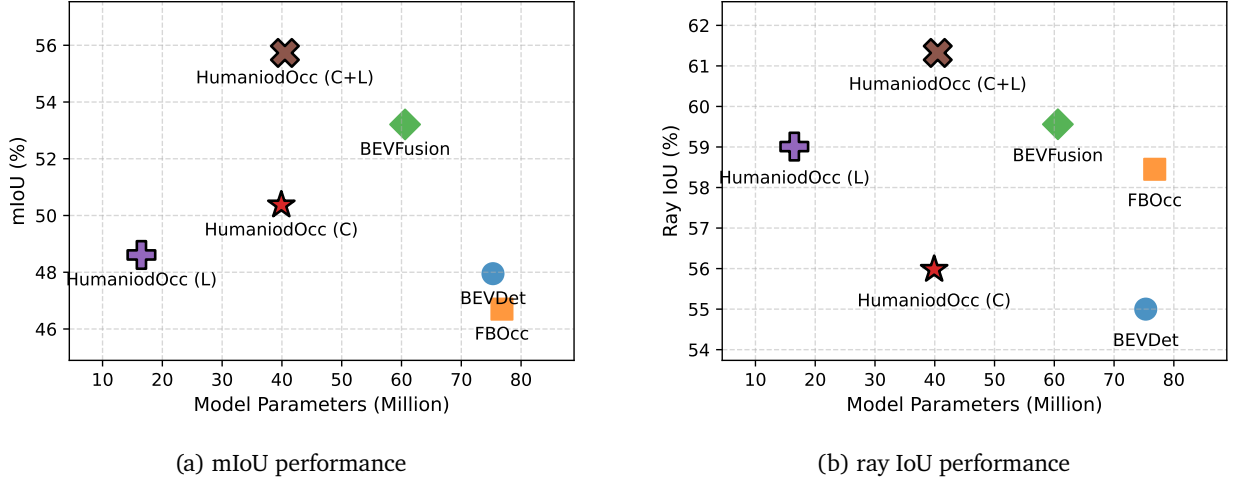


Figure 7: **Comparisons of the mIoU and ray IoU of various 3D occupancy prediction methods.** Our model significantly outperforms other methods in both mIoU and ray IoU with the smallest number of parameters.

Table 3: **Comparisons of the mIoU and ray IoU of the proposed distortion processing methods.** The method we finally implemented avoids undistorting the original image to retain the most effective pixels of the image. Sampling by distorted projection ensures that the performance is not affected.

Method	Input images	Projection	mIoU	rayIoU
HumaniodOcc	raw	pinhole	46.23	52.54
HumaniodOcc	undistorted	pinhole	47.41	53.46
HumaniodOcc	raw	distort	47.92	53.97

modality inputs on occupancy prediction performance.

**Ablation of distortion** As mentioned in Sec. 3.4, we evaluate three distortion handling strategies for images in Tab. 3. Using raw distorted images directly achieves 46.23 mIoU but introduces geometric errors in BEV space. Traditional image undistortion preprocessing improves accuracy to 47.41 mIoU but incurs prohibitive latency for robotic deployment scenarios. Our proposed distortion-aware projection method demonstrates superior performance of 47.92 mIoU by encoding lens distortion parameters directly into view transformation. Notably, for this comparison we use data with precisely calibrated projection alignment to reveal the distinct characteristics of each distortion handling method in terms of geometric accuracy.

**Ablation of temporal** We conduct an ablation study on the number of aggregated frames in Tab. 4. The temporal fusion module effectively aggregates historical frame features, which yields significant improvements over single frame model, primarily by enhancing motion awareness and occlusion reasoning. However, incorporating more than one historical frame leads to performance degradation, which we attribute to the pose errors accumulation during robotic motions.

Table 4: **Comparisons of the mIoU and ray IoU of the different number of temporal frames.** Using 2 frames of temporal information can achieve the best model performance.

Method	Frames	mIoU	rayIoU
HumaniodOcc	1	52.79	60.49
HumaniodOcc	2	55.73	61.32
HumaniodOcc	3	55.11	60.53
HumaniodOcc	4	54.3	60.63

Table 5: **Comparisons of the mIoU and ray IoU of the different number of modalities.** The model can obtain good occupancy prediction results when using only LiDAR information, but the semantic differentiation is relatively poor. The semantic results of the model using only camera information are better than using only LiDAR, but the occupancy prediction is relatively poor due to the lack of clear geometric information in the image. Our fusion method retains the LiDAR geometric information and obtains rich semantic information from the camera.

Method	Modal	mIoU	rayIoU
HumaniodOcc	C	50.37	55.98
HumaniodOcc	L	48.61	59.01
HumaniodOcc	C+L	55.73	61.32

**Ablation of modalities** Tab. 5 demonstrates the expected performance hierarchy: camera-only (50.37 mIoU), LiDAR-only (48.61 mIoU), and multi-modal fusion (55.73 mIoU). The improvement of fusion model confirms the complementary strengths of both modalities - LiDAR provides precise depth while cameras offer rich semantics. This validates our architecture’s effective cross-modal fusion capability.

## 5. Conclusion and Future Work

In this work, we introduced Humanoid Occupancy, a unified multimodal perception system tailored for humanoid robots. By leveraging occupancy-based representations, our framework enables rich semantic and geometric understanding of complex environments, and addresses key challenges in sensor layout, data annotation, and multimodal fusion. We constructed the first panoramic occupancy dataset for humanoid robots and demonstrated the effectiveness of the system on the Tienkung platform in real-world scenarios. Looking ahead, we recognize that omnidirectional perception and mapping are crucial for the future of remote robot control. We plan to utilize state-of-the-art computer vision reconstruction techniques in combination with occupancy representations to build more efficient omnidirectional perception systems (Fig. 8 and Fig. 9), further expanding the dataset, refining temporal fusion strategies, and exploring broader deployment across diverse humanoid platforms to advance robust and standardized visual perception in robotics.

## Acknowledgements

We sincerely thank Hongdi Li and Dengke Shang for their outstanding support and guidance as project managers during the development of this work, and appreciate the dedicated efforts of the X-Humanoid engineering team (Junjie Hu, Qiong Wu, Jian Xiao, Chenghao Sun, Hengle Ren, Wenjing Deng, Hao Yang, Shuo Xu, Dawei Gao, Fachao Zhang, Yuqing Dong) in driving the progress of the Humanoid Occupancy exploratory project. We are especially grateful to the leadership team (Youjun Xiong, Chunzhi Li and Jian Tang) of X-Humanoid for their strategic vision, continuous encouragement, and strong support throughout the project. We also thank Ruiying Zhang for helping with the production of Fig. 6, and Gabriele Meilikhov for his contributions to the visual presentation.

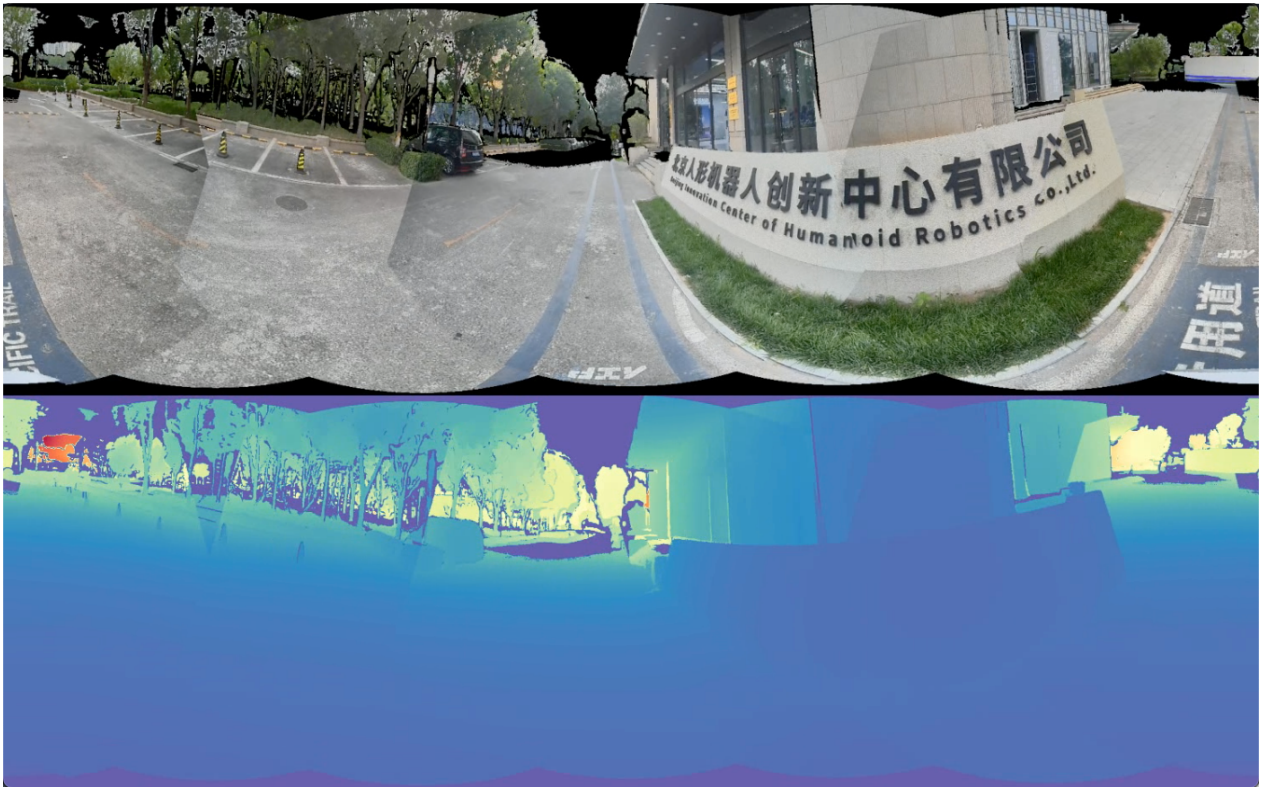


Figure 8: Panoramic stitching of 6 cameras and densified visualization of LiDAR point clouds.

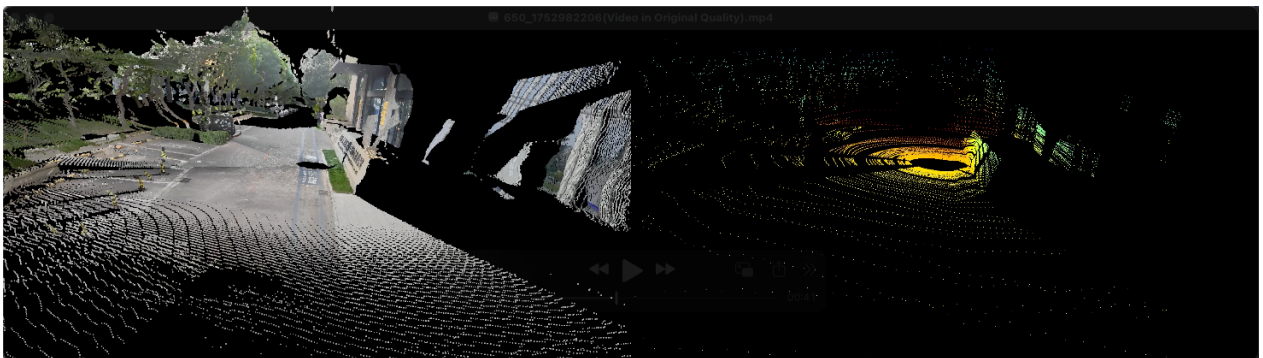


Figure 9: 3D scene visualization and LiDAR point clouds visualization.



## References

- [1] Evan Ackerman. Video friday: Atlas robot sees the world. URL <https://spectrum.ieee.org/video-friday-atlas-robot-sees-world>. 3
- [2] AgilityRobotics. agilityrobotics website. URL <https://www.agilityrobotics.com/>. 3
- [3] Qingwei Ben, Feiyu Jia, Jia Zeng, Juntong Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025. 3
- [4] Andrea Bonci, Pangcheng David Cen Cheng, Marina Indri, Giacomo Nabissi, and Fiorella Sibona. Human-robot perception in industrial environments: A survey. *Sensors*, 21(5):1571, 2021. 2
- [5] 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. 5
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [7] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024. 3
- [8] Carl F DiSalvo, Francine Gemperle, Jodi Forlizzi, and Sara Kiesler. All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 321–326, 2002. 2
- [9] Helei Duan, Bikram Pandit, Mohitvishnu S Gadde, Bart Van Marum, Jeremy Dao, Chanh Kim, and Alan Fern. Learning vision-based bipedal locomotion for challenging terrain. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 56–62. IEEE, 2024. 3
- [10] Alberto Elfes. Occupancy grids: A stochastic spatial representation for active robot perception. *arXiv preprint arXiv:1304.1098*, 2013. 3
- [11] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024. 3
- [12] Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *arXiv preprint arXiv:2501.02116*, 2025. 2
- [13] Junzhe He, Chong Zhang, Fabian Jenelten, Ruben Grandia, Moritz BÄcher, and Marco Hutter. Attention-based map encoding for learning generalized legged locomotion. *arXiv preprint arXiv:2506.09588*, 2025. 3
- [14] Kazuo Hirai, Masato Hirose, Yuji Haikawa, and Toru Takenaka. The development of honda humanoid robot. In *Proceedings. 1998 IEEE international conference on robotics and automation (Cat. No. 98CH36146)*, volume 2, pages 1321–1326. IEEE, 1998. 2
- [15] Masato Hirose and Kenichi Ogawa. Honda humanoid robots development. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1850):11–19, 2007. 3
- [16] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 9

- [17] Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 10
- [18] 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*, pages 9223–9232, 2023. 5
- [19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4
- [20] Daehwa Kim, Mario Srouji, Chen Chen, and Jian Zhang. Armor: Egocentric perception for humanoid robot collision avoidance and motion planning. *arXiv preprint arXiv:2412.00396*, 2024. 3
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3
- [22] 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, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12697–12705. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01298. URL [http://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Lang\\_PointPillars\\_Fast\\_Encoders\\_for\\_Object\\_Detection\\_From\\_Point\\_Clouds\\_CVPR\\_2019\\_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Lang_PointPillars_Fast_Encoders_for_Object_Detection_From_Point_Clouds_CVPR_2019_paper.html). 8
- [23] 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*, pages 9087–9098, 2023. 5
- [24] Yingwei Li, Adams Wei Yu, Tianjian Meng, Benjamin Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan L. Yuille, and Mingxing Tan. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17161–17170. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01667. URL <https://doi.org/10.1109/CVPR52688.2022.01667>. 8
- [25] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 5, 10
- [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*, 35:10421–10434, 2022. 5
- [27] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.106. URL <https://doi.org/10.1109/CVPR.2017.106>. 9
- [28] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 10
- [29] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 5, 10

- [30] Junfeng Long, Junli Ren, Moji Shi, Zirui Wang, Tao Huang, Ping Luo, and Jiangmiao Pang. Learning humanoid locomotion with perceptive internal model, 2024. URL <https://arxiv.org/abs/2411.14386>. 3
- [31] Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. *arXiv preprint arXiv:2412.07773*, 2024. 3
- [32] Shixin Luo, Songbo Li, Ruiqi Yu, Zhicheng Wang, Jun Wu, and Qiuguo Zhu. Pie: Parkour with implicit-explicit learning framework for legged robots. *IEEE Robotics and Automation Letters*, 2024. 4
- [33] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5
- [34] Steve Macenski, Alberto Soragna, Michael Carroll, and Zhenpeng Ge. Impact of ros 2 node composition in robotic systems. *IEEE Robotics and Automation Letters*, 8(7):3996–4003, 2023. 4
- [35] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. Robot operating system 2: Design, architecture, and uses in the wild. *Science robotics*, 7(66):eabm6074, 2022. 4
- [36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4
- [37] Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. Occfusion: Multi-sensor fusion framework for 3d semantic occupancy prediction. *IEEE Transactions on Intelligent Vehicles*, 2024. 5
- [38] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica Hodgins, and Sara Kiesler. Perceptions of asimo: an exploration on co-operation and competition with humans and humanoid robots. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 351–352, 2006. 3
- [39] Daojie Peng, Jiahang Cao, Qiang Zhang, and Jun Ma. Lovon: Legged open-vocabulary object navigator. *arXiv preprint arXiv:2507.06747*, 2025. 3
- [40] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020. 2
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [42] Arindam Roychoudhury, Shahram Khorshidi, Subham Agrawal, and Maren Bennewitz. Perception for humanoid robots. *Current Robotics Reports*, 4(4):127–140, 2023. 2
- [43] Nikita Rudin, Junzhe He, Joshua Aurand, and Marco Hutter. Parkour in the wild: Learning a general and extensible agile locomotion policy using multi-expert distillation and rl fine-tuning. *arXiv preprint arXiv:2505.11164*, 2025. 3
- [44] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *2011 IEEE international conference on robotics and automation*, pages 1–4. IEEE, 2011. 3
- [45] Satoshi Shigemitsu. Asimo and humanoid robot research at honda. In *Humanoid robotics: A reference*, pages 1–36. Springer, 2017. 3

- [46] Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. *arXiv preprint arXiv:2306.02851*, 2023. 4
- [47] Stanford Artificial Intelligence Laboratory et al. Robotic operating system. URL <https://www.ros.org>. 4
- [48] Jingkai Sun, Qiang Zhang, Gang Han, Wen Zhao, Zhe Yong, Yan He, Jiaxu Wang, Jiahang Cao, Yijie Guo, and Renjing Xu. Trinity: A modular humanoid robot ai system. *arXiv preprint arXiv:2503.08338*, 2025. 3
- [49] Wandong Sun, Baoshi Cao, Long Chen, Yongbo Su, Yang Liu, Zongwu Xie, and Hong Liu. Learning perceptive humanoid locomotion over challenging terrain. *arXiv preprint arXiv:2503.00692*, 2025. 3
- [50] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15035–15044, 2024. 5
- [51] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36:64318–64330, 2023. 5
- [52] Hao Wang, Xiaobao Wei, Xiaolan Zhang, Jianing Li, Chengyu Bai, Ying Li, Ming Lu, Wenzhao Zheng, and Shanghang Zhang. Embodiedocc++: Boosting embodied 3d occupancy prediction with plane regularization and uncertainty sampler, 2025. URL <https://arxiv.org/abs/2504.09540>. 5
- [53] Jiaxu Wang, Qiang Zhang, Jingkai Sun, Jiahang Cao, Gang Han, Wen Zhao, Weining Zhang, Yecheng Shao, Yijie Guo, and Renjing Xu. Reinforcement learning with generalizable gaussian splatting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 435–441. IEEE, 2024. 4
- [54] Jiaxu Wang, Ziyi Zhang, Qiang Zhang, Jia Li, Jingkai Sun, Mingyuan Sun, Junhao He, and Renjing Xu. Query-based semantic gaussian field for scene representation in reinforcement learning. *arXiv preprint arXiv:2406.02370*, 2024. 4
- [55] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 10
- [56] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 5
- [57] Zifan Wang, Teli Ma, Yufei Jia, Xun Yang, Jiaming Zhou, Wenlong Ouyang, Qiang Zhang, and Junwei Liang. Omni-perception: Omnidirectional collision avoidance for legged locomotion in dynamic environments. *arXiv preprint arXiv:2505.19214*, 2025. 3, 4
- [58] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Embodiedocc: Embodied 3d occupancy prediction for vision-based online scene understanding. *arXiv preprint arXiv:2412.04380*, 2024. 5
- [59] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 5, 9
- [60] Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on robot learning*, pages 284–301. PMLR, 2023. 4



- [61] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024. 3
- [62] Ji Zhang, Yiran Ding, and Zixin Liu. Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction. In *Proceedings of the Asian Conference on Computer Vision*, pages 3587–3604, 2024. 5
- [63] Qiang Zhang, Gang Han, Jingkai Sun, Wen Zhao, Chenghao Sun, Jiahang Cao, Jiaxu Wang, Yijie Guo, and Renjing Xu. Distillation-ppo: A novel two-stage reinforcement learning framework for humanoid robot perceptive locomotion. *arXiv preprint arXiv:2503.08299*, 2025. 3
- [64] Qiang Zhang, Zhang Zhang, Wei Cui, Jingkai Sun, Jiahang Cao, Yijie Guo, Gang Han, Wen Zhao, Jiaxu Wang, Chenghao Sun, et al. Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots. *arXiv preprint arXiv:2503.09010*, 2025. 3, 5
- [65] Yang Zhang, Zhanxiang Cao, Buqing Nie, Haoyang Li, and Yue Gao. Learning robust motion skills via critical adversarial attacks for humanoid robots. *arXiv preprint arXiv:2507.08303*, 2025. 3
- [66] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 5
- [67] Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Hengle Ren, Renjing Xu, and Jian Tang. Roboocc: Enhancing the geometric and semantic scene understanding for robots. *arXiv preprint arXiv:2504.14604*, 2025. 4
- [68] Zhang Zhang, Qiang Zhang, Wei Cui, Shuai Shi, Yijie Guo, Gang Han, Wen Zhao, Jingkai Sun, Jiahang Cao, Jiaxu Wang, et al. Occupancy world model for robots. *arXiv preprint arXiv:2505.05512*, 2025. 4
- [69] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 4
- [70] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=gZ9hCDWe6ke>. 9
- [71] Ziwen Zhuang, Shenzhe Yao, and Hang Zhao. Humanoid parkour learning. *arXiv preprint arXiv:2406.10759*, 2024. 3