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

Wei Cui*[1], Haoyu Wang*[2], Jiaru Zhong[1], Wenkang Qin[2], Yijie Guo[1], Gang Han[1], Wen Zhao[1], Jiahang Cao[1], Zhang Zhang[1], Jingkai Sun[1], Pihai Sun[1], Shuai Shi[1], Botuo Jiang[1], Jiahao Ma[1], Jiaxu Wang[1], Hao Cheng[1], Zhichao Liu[2], Yang Wang[2], Zheng Zhu[2], Guan Huang[2], Lingfeng Zhang[3], Jun Ma[4], Junwei Liang[4], Renjing Xu[4], Jian Tang†[1] and Qiang Zhang†Δ[1]

***Abstract*— As the most promising general-purpose robots, humanoid robots have witnessed rapid development. The visual perception module serves as a crucial component within their system, providing essential information for key tasks such as mobile manipulation. Therefore, this paper proposes a perception framework centered on semantic occupancy, termed Humanoid Occupancy, enhancing perception capabilities by optimized scene representation and efficient data fusion. Specifically, we design a sensor arrangement scheme, develop the first panoramic occupancy dataset for humanoid robots, and propose a lightweight multi-modal temporal feature fusion network which can present the dynamic environments precisely. Furthermore, this perception method has been deployed on the Tienkung humanoid robot platform, demonstrating its superior environmental perception and navigation performance in changing environments.**

I. Introduction

Humanoid robots are expected to perform autonomous navigation, locomotion, and manipulation in dynamic environments, assisting humans in work and daily life through the realization of mobile manipulation. Since these advanced tasks rely on precise and comprehensive environmental perception, this paper focuses on enhancing the environmental perception capabilities of humanoid robots.

Existing perception systems for humanoid robots incorporate multiple sensors, including monocular or stereo cameras [1], RGB-D cameras [2], LiDAR [3], and panoramic cameras [4]. They utilize deep neural networks for feature extraction, multimodal data fusion, and other processes to generate perception results. With the advancement of network architectures, visual perception systems for humanoid robots are evolving toward multimodal fusion [5], optimized spatial representation [6], and end-to-end decision-making [7]. Therefore, how to efficiently arrange sensors within limited space and achieve effective data fusion and representation has become a core issue for improving system performance.

Inspired by progress in autonomous driving and robot navigation [8], we introduce semantic occupancy [9], which simultaneously encodes occupancy status and semantic labels, into humanoid robot perception systems. Semantic occupancy offers two key advantages. First, occupancy can directly encode the occupancy status of each spatial unit in the environment, 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), which typically focus only on ground projections and fail to reflect vertical structures and high-level semantic information. Second, occupancy representation 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 or meshes, occupancy offers greater generality and extensibility in data structure, semantic annotation, and downstream task interfaces. 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.

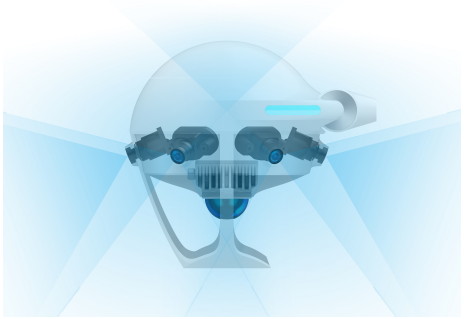

Fig. 1. **Sensor Configuration for Humanoid Occupancy.**

However, using off-the-shelf occupancy prediction networks [10]–[13] and datasets [9], [14] cannot solve the problems specific to humanoid robots, as humanoid robots require high-resolution occupancy maps within short ranges and must address frequent occlusions and articulated motions. Furthermore, although recent pioneering methods [6] have attempted to construct occupancy perception approaches for robots, they suffer from limitations in multimodal fusion, real-time inference, and dataset construction.

Based on the above observations, this paper proposes the **Humanoid Occupancy** system, a full-stack perception solution dedicated to humanoid robots, encompassing hardware design, dataset construction, and a multimodal occupancy prediction network. First, we design a sensor layout strategy tailored for humanoid robots, effectively mitigating

*Equal Contributors, †Corresponding Authors, ΔProject and Technical Leader. Email: jony.zhang@x-humanoid.com
[1]X-Humanoid, [2]GigaAI, [3]THU, [4]HKUST(GZ)
Project Page: https://humanoid-occupancy.github.io/

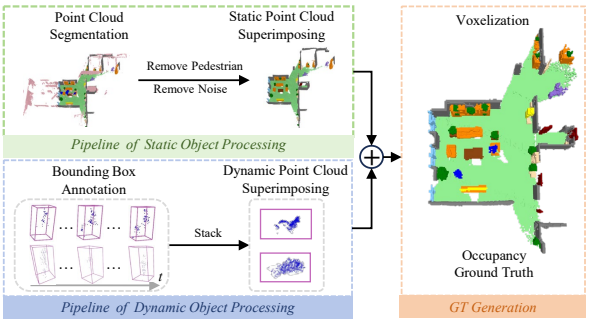

Fig. 2. **Dynamic-Static Decoupled Annotation Pipeline.**

perception blind spots. For the dataset, we design a low-cost, high-efficiency collection scheme and adopt a dynamic-static decoupling annotation method to obtain ground-truth occupancy labels. Finally, we develop a multimodal temporal occupancy prediction network based on LiDAR point clouds and camera images, which achieves excellent perception performance while ensuring lightweight deployment. More importantly, we have deployed it on the Tienkung robot, demonstrating real-time navigation based on the occupancy prediction in dynamic environments, which validates the application prospects of our proposed Humanoid Occupancy.

II. Method

In this section, we elaborate on the proposed Humanoid Occupancy, including subsection II-A on sensor configuration, subsection II-B on dataset construction, and subsection II-C on the occupancy prediction network, HumanoidOcc.

A. Sensor Configuration

Considering the perception requirements of different tasks and the constraints of the robot's structure, we design a sensor suite comprising 6 RGB cameras and 1 LiDAR, which are mounted on the robot's head to achieve a 360° coverage range, as illustrated in Fig. 1.

B. Dataset Construction

1) Data Acquisition: Given the high cost and difficulty of directly collecting data using humanoid robots, we design a wearable device that maintains an identical sensor configuration to that on the robot. The data collector's height is approximately 160 cm, comparable to that of the Tienkung robot. Additionally, to prevent head shaking by the collector, which would cause loss of horizontal stability, we provide the collector with a neck stabilizer. To enhance the diversity of the dataset, we arrange multiple scenarios, including household, industrial, and outdoor environments. The collector walked slowly in these scenarios while recording data from all sensors and other essential information.

2) Annotation: To ensure the accuracy of annotations, we adopt a dynamic-static decoupled annotation workflow as illustrated in Fig. 2. For dynamic objects, including pedestrians, cyclists, and vehicles, we describe them frame-by-frame using bounding boxes. Note that in indoor scenarios, pedestrians may exhibit diverse postures (e.g., squatting), which cannot be easily annotated with bounding boxes. Thus, we implement special process for pedestrians: ordinary postures are annotated solely with bounding boxes, while special postures are processed using the annotation method for static background. For static background, we first remove the already annotated dynamic objects frame-by-frame, then superimpose point clouds from multiple frames to form a dense point cloud. Point-wise semantic segmentation is then applied to the point cloud by projecting it onto images. All annotation processes were conducted by professional annotators and subjected to quality checks to ensure annotation accuracy. After completing the annotations, we first align the superimposed static background point cloud frame-by-frame to the robot's egocentric coordinate system based on motion information. Subsequently, we splice the point clouds corresponding to dynamic objects into the point cloud according to their poses. Finally, the resulting superimposed point cloud is directly voxelized, and the final ground-truth semantic occupancy is obtained based on the labels.

C. HumaniodOcc Network

As illustrated in Fig. 3, our proposed HumanoidOcc occupancy prediction network takes LiDAR point clouds and surround-view images as inputs. These inputs are processed by a point cloud encoder and an image encoder respectively to generate modality-specific features. Subsequently, a multimodal fusion transformer efficiently fuses the two modal features in the Bird's Eye View (BEV) space. To further enhance perceptual capability, historical BEV features are incorporated to generate the final scene representation, which is used to produce the occupancy prediction results. Notably, due to the robot's pitch and roll motions, we choose to construct BEV features in a gravity-aligned egocentric reference frame to align with the assumptions of BEV.

Image Feature Encoder. A network consisting of residual convolutions and a feature pyramid is employed to extract features from images, with parameter sharing across all 6 cameras. To reduce system latency, we directly process distorted images, while correspondingly adjusting the projection in the multimodal feature fusion network. Ablation study III-C.1 shows that this choice achieves optimal performance.

Point Cloud Feature Encoder. For point clouds, the classic PointPillar [15] is used to construct BEV features. Specifically, the spatial region containing the point cloud is first uniformly divided into a set of pillars. To reduce computational complexity, only non-empty pillars (those containing points) are extracted for further processing. The coordinates, attributes, and other properties of points within each pillar are encoded into pillar-specific features. These features are then scattered into the BEV space according to the original coordinates of the pillars, generating a sparse BEV feature map. Finally, a 2D convolution backbone captures multi-scale spatial relationships to form dense BEV features.

Multimodal Feature Fusion. Inspired by DeepFusion [16], we adopt a cross-attention mechanism for multimodal fusion, where point cloud features serve as queries and image features act as keys and values. Specifically, deformable attention [17] is utilized to enable fusion: point cloud feature coordinates are projected onto the image plane

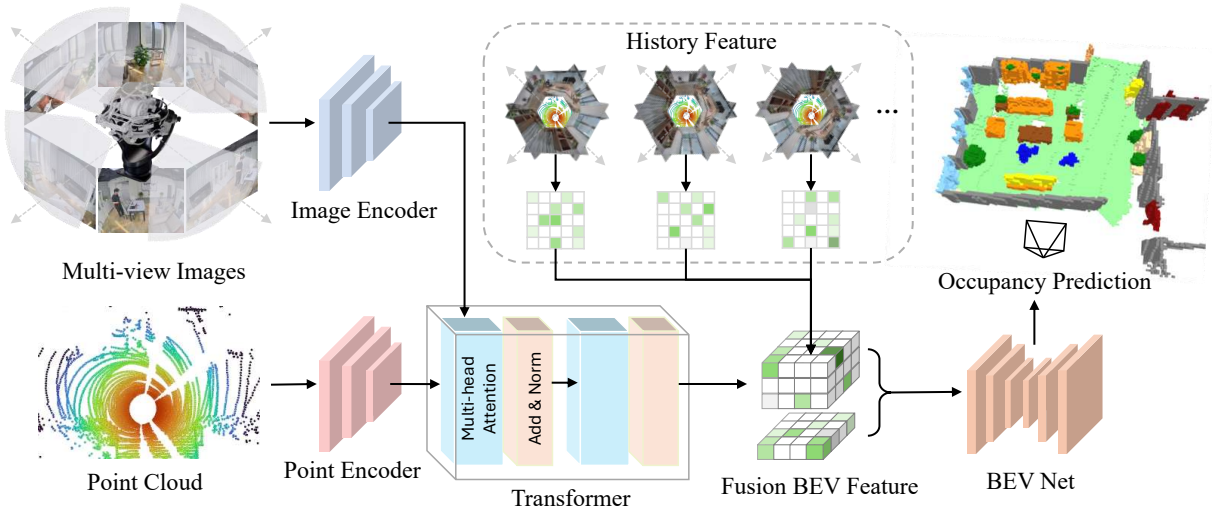

Fig. 3. **Overview of Our Proposed HumaniodOcc.** The proposed network processes image and LiDAR inputs through separate encoders to extract features. The multimodal features are then fused via cross-attention, enabling adaptive interaction between visual and geometric cues. The fused features undergo temporal fusion to aggregate sequential information. Finally, the network predicts the 3D occupancy using a prediction head.

using a pinhole camera distortion projection model, nearby features are sampled, and cross-attention interaction with point cloud features is performed to generate the final fused features, which integrate both geometric information from point clouds and semantic information from images.

Historical Feature Fusion. In perception tasks, fusing historical features helps the model capture temporal information and achieve more accurate environmental representation, which is crucial for fine-grained occupancy prediction. Following BEVDet4D [18], we maintain a BEV feature queue that dynamically updates its contents at each timestamp. During fusion, historical features are first aligned to the current coordinate system using bilinear interpolation based on ego-motion between two timestamps. They are then concatenated with current BEV features along the channel dimension. Finally, a ResNet-FPN hybrid architecture fully fuses historical features across multiple resolutions to generate the enhanced final BEV features.

Prediction Head. To enable efficient occupancy prediction, BEV features are lifted into the 3D voxel space using the channel-to-height transformation proposed by FlashOcc [19], followed by a 3D convolution block to predict the final results, including occupancy status and semantic categories.

III. Experiment

A. Dataset and Implementation

Dataset. The dataset comprises 200 sequences, each containing 200 consecutive frames to ensure rich contextual information. The dataset is split into training and validation sets at a ratio of 9:1. Following existing methodologies [20], we use mIoU and rayIoU to evaluate occupancy prediction accuracy, while also counting the number of model parameters to quantify model size.

Training Settings. The perception range is set to [-10m, 10m] along the X and Y axes, and [-1.5m, 0.9m] along the Z axis in the ego coordinate system. The predicted output size is $200 \times 200 \times 24$, with each voxel sized at [0.1m, 0.1m, 0.1m]. Input images are cropped to a resolution of 960×768. The AdamW optimizer is used with an initial learning rate of 2e-4 and a cosine annealing scheduling strategy. Training is conducted with a batch size of 4 for 20 epochs using 8 NVIDIA A100 GPUs.

B. Main Results

We established a benchmark on our dataset, including BEVDet [11], FBOcc [10], and BEVFusion [12], under both single-frame and multi-frame settings. The experimental results are presented in the Table I. Compared to baselines, our model achieves superior performance with the smallest number of parameters. When incorporating 1 frame of historical features, HumanoidOcc exhibits a significant performance improvement, with mIoU and rayIoU increasing by 2.94 and 0.83 respectively, which demonstrates the important role of temporal information.

C. Ablation Study

1) Ablation of Distortion: The process of distorted images significantly impacts both the image branch and the multimodal fusion network. Therefore, as shown in Table II, we compare the performance under different input image conditions and projection methods. Using raw distorted images as input with a standard pinhole camera projection achieves 46.23 mIoU and 52.54 rayIoU. When inputting undistorted images, mIoU and rayIoU improve by 1.18 and 0.92 respectively, which can be attributed to more accurate geometric correspondence between point clouds and images. The distortion-aware projection method we adopt achieves the optimal performance, further improving mIoU and rayIoU by 0.51 each while eliminating the need for an additional distortion correction step.

2) Effect of Temporal Information: To investigate the impact of temporal information, we gradually increase the number of frames in the historical feature fusion module from 0 to 3, with the results presented in Table III. The findings indicate that fusing historical features effectively improves occupancy prediction accuracy, primarily attributed to enhanced motion perception and occlusion reasoning.

TABLE I

3D SEMANTIC OCCUPANCY PREDICTION PERFORMANCE ON OUR DATASET

Method	Settings		Params	Metrics		Class Metrics											
	Modality	Frames		mIoU	rayIoU	pedestrian	robot	chair	table	floor	wall	window	door	plant	appliance	furniture	objects
BEVDet [11]	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 [10]	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 [12]	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
HumaniodOcc (Ours)	C+L	1	40.5M	52.79	60.49	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 [11]	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 [10]	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 [12]	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
HumaniodOcc (Ours)	C+L	2	40.5M	55.73	61.32	50.85	56.16	40.49	45.49	65.60	64.76	44.07	72.77	65.58	63.01	64.50	35.48

TABLE II

PERFORMANCE OF DIFFERENT DISTORTION PROCESSING METHODS

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

TABLE III

EFFECT OF DIFFERENT NUMBER OF HISTORICAL FRAMES

Method	Historical Frames	mIoU	rayIoU
HumaniodOcc	0	52.79	60.49
HumaniodOcc	1	55.73	61.32
HumaniodOcc	2	55.11	60.53
HumaniodOcc	3	54.3	60.63

However, fusing more historical frames leads to performance degradation, which may be caused by the accumulation of pose errors due to the complex motion of the robot.

3) Ablation of Modalities: We conducted modal ablation, and the results are shown in Table IV. It can be observed that multimodal fusion indeed brings performance improvements, with mIoU increasing by 5.36 and 7.12 compared to camera-only and lidar-only methods, respectively. This demonstrates the complementarity of images and point clouds in occupancy prediction tasks, where point clouds provide geometric information and images offer semantic features.

TABLE IV

PERFORMANCE OF DIFFERENT MODALITIES

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

IV. CONCLUSION

In this paper, we propose Humanoid Occupancy, a multi-modal perception system specifically designed for humanoid robots, enhancing the understanding of semantic and geometric structures of the surrounding environment. To integrate the semantic occupancy representation, this framework addresses the challenges in sensor arrangement, dataset construction, and occupancy prediction. Humanoid Occupancy is expected to advance humanoid robot perception research and boost their practical application.

REFERENCES

[1] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," *arXiv preprint arXiv:2502.13013*, 2025.

[2] Q. Zhang, G. Han, J. Sun, W. Zhao, C. Sun, J. Cao, J. Wang, Y. Guo, and R. Xu, "Distillation-ppo: A novel two-stage reinforcement learning framework for humanoid robot perceptive locomotion," *arXiv preprint arXiv:2503.08299*, 2025.

[3] Z. Wang, T. Ma, Y. Jia, X. Yang, J. Zhou, W. Ouyang, Q. Zhang, and J. Liang, "Omni-perception: Omnidirectional collision avoidance for legged locomotion in dynamic environments," *arXiv preprint arXiv:2505.19214*, 2025.

[4] Q. Zhang, Z. Zhang, W. Cui, J. Sun, J. Cao, Y. Guo, G. Han, W. Zhao, J. Wang, C. Sun, *et al.*, "Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots," *arXiv preprint arXiv:2503.09010*, 2025.

[5] N. Rudin, J. He, J. Aurand, and M. 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.

[6] Z. Zhang, Q. Zhang, W. Cui, S. Shi, Y. Guo, G. Han, W. Zhao, H. Ren, R. Xu, and J. Tang, "Roboocc: Enhancing the geometric and semantic scene understanding for robots," *arXiv preprint arXiv:2504.14604*, 2025.

[7] S. Luo, S. Li, R. Yu, Z. Wang, J. Wu, and Q. Zhu, "Pie: Parkour with implicit-explicit learning framework for legged robots," *IEEE Robotics and Automation Letters*, 2024.

[8] D. Peng, J. Cao, Q. Zhang, and J. Ma, "Lovon: Legged open-vocabulary object navigator," *arXiv preprint arXiv:2507.06747*, 2025.

[9] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 318–64 330, 2023.

[10] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.

[11] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[12] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022.

[13] J. Zhong, J. Wang, J. Xu, X. Li, Z. Nie, and H. Yu, "Cooptrack: Exploring end-to-end learning for efficient cooperative sequential perception," *arXiv preprint arXiv:2507.19239*, 2025.

[14] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 850–17 859.

[15] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[16] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 182–17 191.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[18] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[19] Z. Yu, C. Shu, J. Deng, K. Lu, Z. Liu, J. Yu, D. Yang, H. Li, and Y. Chen, "Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin," *arXiv preprint arXiv:2311.12058*, 2023.

[20] H. Liu, Y. Chen, H. Wang, Z. Yang, T. Li, J. Zeng, L. Chen, H. Li, and L. Wang, "Fully sparse 3d occupancy prediction," in *European Conference on Computer Vision*. Springer, 2024, pp. 54–71.